EVOLUTION OF PROTEIN ABUNDANCES IN EUKARYOTES

DISSERTATION ZUR ERLANGUNG

DER NATURWISSENSCHAFTLICHEN DOKTORWÜRDE (DR. SC. NAT.)

VORGELEGT DER

MATHEMATISCH-NATURWISSENSCHAFTLICHEN FAKULTÄT

DER

UNIVERSITÄT ZÜRICH

VON

MANUEL WEISS

PROMOTIONSKOMITEE

PROF. MICHAEL O. HENGARTNER

(VORSITZ, LEITER DER DISSERTATION)

PROF. CHRISTIAN VON MERING

PROF. FABIO PIANO

DR. PEER BORK

ZÜRICH 2010

Table of contents

1 Zusammenfassung

Mit modernen Massenspektrometern können hunderte von Proteinen in einer komplexen biologischen Probe in wenigen Stunden identifiziert werden. Eine weit verbreitete Methode, Shotgun Proteomics, erlaubt auch die Bestimmung von semi-quantitativer Information über die Menge der gemessenen Proteine. Wir verwendeten dieses sogenannte Spectral Counting auf mehreren öffentlich verfügbaren grossen Datensätzen von fünf verschiedenen Arten. Wir konnten zeigen, dass die Proteinmengen bemerkenswert gut konserviert sind, zumindest für das hier betrachtete Kernproteom, das hauptsächlich aus uralten, konservierten Proteinen mit grundlegenden Funktionen besteht. Wir arbeiten nun an einer Online-Datenbank von Proteinabundanzen, um diese Informationen der wissenschaftlichen Gemeinschaft zur Verfügung zu stellen.

2 Summary

With modern mass spectrometers, hundreds of proteins in a complex biological sample can be identified within a few hours. One common method, shotgun proteomics, also allows the inference of semi-quantitative abundance information about the measured proteins. We used this so-called spectral counting on several publicly available large data sets from five different species. We could show that the protein abundances are remarkably well conserved, at least for the considered core proteome that consists mostly of ancient, conserved proteins with housekeeping functions. We are now developing an online database of protein abundances to make this information available to the scientific community.

3 Introduction

3.1 *C. elegans* as a model organism

The nematode *C. elegans* is a popular model organism, as it has a rapid life cycle, is easy to breed and can be frozen and thawed, allowing long-term storage. It has a fixed number of cells in the adult and is transparent, so many processes can be studied *in vivo* under the microscope (Wood 1988).

For example, *C. elegans* is a good model to study the mechanisms related to cell death, as some of the cells undergo programmed cell death (apoptosis) in a predictable manner. Most of the responsible genes have orthologs in human and are therefore important for cancer research (Kinchen and Hengartner 2005; Lettre and Hengartner 2006).

In *C. elegans*, it is also relatively easy to knock down genes by RNAi, for example by feeding bacteria that express a certain RNA (Dudley and Goldstein 2005).

3.2 From genomics to proteomics

C. elegans has about 20,000 predicted genes. It was the first multicellular organism with a fully sequenced genome, which was published in 1998 (Hodgkin et al. 1998). Many of the genes have yet not been verified by experiments, so their expression status is unknown.

In the last ten years, whole-genome sequencing has become a lot cheaper and faster and to date, the genomes of almost 100 eukaryotes have been sequenced (EBI 2010).

Nevertheless, the genome is only the blueprint of the organism and does not give any detailed information as to which gene is expressed when and where, and at which amount, if at all. In addition, the genome is static while

the proteome changes under different environmental conditions and biological stages.

With the advent of high-precision mass spectrometers, the measurement of whole proteomes is within reach. Now it is possible to verify and quantify the expression of genes on a large scale. To a certain extent, this was already possible using transcriptome data from microarrays, but these detect mRNA and not the end product, i.e. the protein. This might be enough to verify expression as such but does not give much quantitative information as many different factors can influence the transcript-to-protein ratio, including mRNA degradation, translation efficiency, protein half-life, etc (Greenbaum et al. 2003).

The use of high-throughput methods has already led to the generation of a considerable amount of data and – just as in genomics – this amount is likely to grow exponentially in the next few years. This makes it extremely important to use open standards and common procedures to store, share and mine these data. At the same time, it has become very important to develop bioinformatics tools to handle and analyze all this information, especially considering the need for advanced statistics to control false discovery rates and to prevent misleading conclusions further downstream (Deutsch, Lam, and Ruedi Aebersold 2008; Reiter et al. 2009).

3.3 From protein catalogues to quantification

So far, the purpose of these large-scale proteomics efforts has been to identify as many proteins as possible for a given organism, that is to compile a catalogue of proteins amenable to measurement by mass spectrometry. Such a list already gives many valuable biological insights and also allows a compilation of so-called proteotypic peptides (PTPs, also see 3.3.1), i.e. peptides that are known to be observable in the mass spectrometer and to

uniquely identify a specific protein, to be used for more specific measurements later.

Beyond simple identification, information about the actual abundance of a protein in a given sample opens the door to a whole new world of biological discoveries.

Before high-throughput mass spectrometry was available, methods such as western blotting, protein electroblotting and Edman sequencing were used. These were low-throughput methods mostly used to identify and/or quantify single proteins. By mass spectrometry, on the other hand, hundreds of proteins can be identified in one measurement, allowing systems-level analyses (Gevaert and Vandekerckhove 2000).

3.3.1 Mass spectrometry-based proteomics

For shotgun proteomics, one of the most common methods, a complex protein sample (optionally already fractionated) is digested with trypsin, and then a variety of separation methods can be used to reduce the complexity of the peptide mixture before it is loaded into a liquid chromatography column. The peptides are then measured on a connected tandem mass spectrometer and with the help of spectrum-to-sequence-matching algorithms (called "search engines") they can be identified and hence the protein inferred. These algorithms compare the recorded spectrum with a huge list of theoretically derived spectra that is compiled by *in-silico* digestion of all protein sequences of a given organism and computation of the possible fragmentation patterns of each of these peptides. This is a computationally very expensive procedure and obviously only allows detection of proteins contained in the list of sequences (Steen and Matthias Mann 2004; Marcotte 2007).

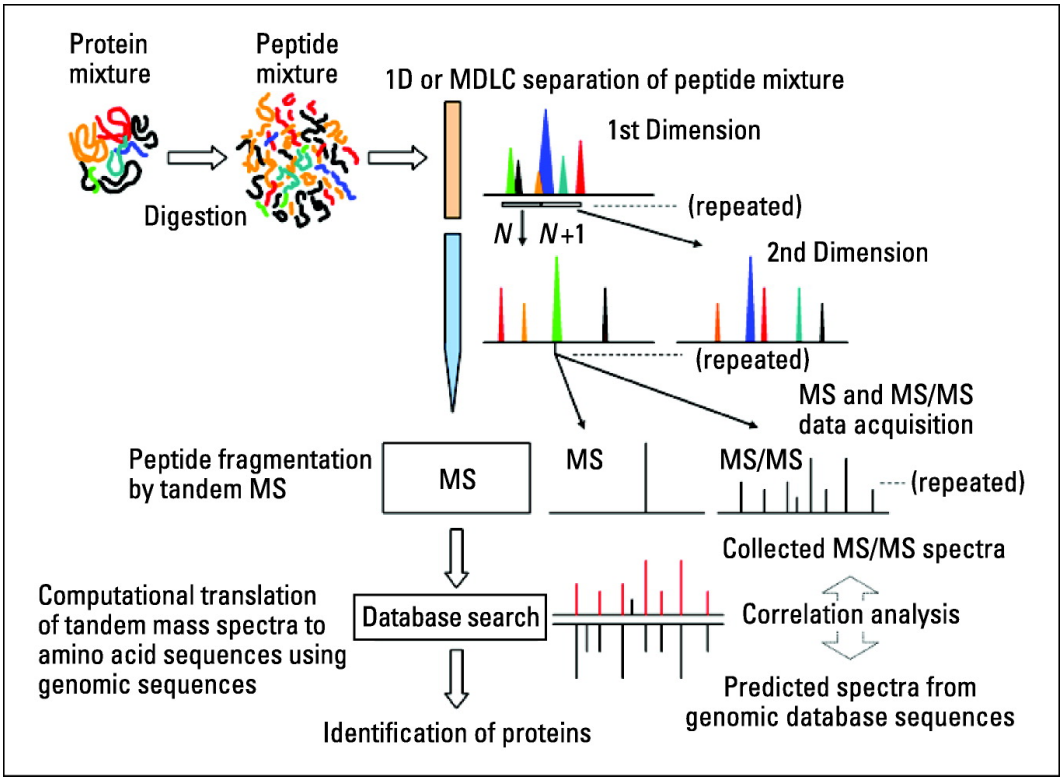


Figure 1: Standard shotgun proteomics workflow (from (Motoyama and Yates 2008))

An alternative is de-novo sequencing, where the peptide sequence is reconstructed straight from the spectrum (Bandeira et al. 2008).

Shotgun proteomics is not a very efficient way of identifying and quantifying specific proteins, as a lot of instrument time is spent either measuring the same abundant proteins again and again or, more generally, measuring proteins the researcher is not interested in. Yet, it is an unbiased approach and allows the identification of novel proteins. The recent development of multiple reaction monitoring however, also called selected reaction monitoring (MRM/SRM), represents a targeted approach where a pre-defined list of peptides is given to the mass spectrometer, enabling hypothesis-driven research. The machine then tries to exclusively look for these peptides, ignoring everything else and therefore increasing the signal-to-noise ratio dramatically (Lange et al. 2008; Prakash et al. 2009).

This also makes the measurement relatively reproducible which is not the case with normal shotgun proteomics where the peptides are selected by chance and technical replicates are known to have relatively little overlap.

The drawback of MRM/SRM is the required peptide inclusion list. It has to be generated by first collecting a large catalogue of proteotypic peptides (PTPs) via regular shotgun proteomics or by using bioinformatics tools to predict these PTPs. Unfortunately, these tools are known to not perform very well.

3.3.2 Spectral counting

Shotgun proteomics does not deliver quantitative information per se, but if the collected spectra are numerous enough, relative abundance information can be inferred using "spectral counting" (Lu et al. 2007). The idea behind this is the fact that abundant proteins tend to generate more spectra and therefore more peptide identifications. Once the peptide counts are normalized for the number of detectable peptides for a certain protein (a parameter that can be learned from the data, see chapter 4 for details), a value can be calculated that reflects the average number of identifications of any amino acid of this protein which in turn is proportional to the relative abundance of the protein in the sample. Again, this works only for very large data sets where most of the proteins have multiple peptide identifications. The actual calculations are explained in more detail in chapter 4 and 5.

As we show in chapter 5, this relatively simple method compares quite well to other approaches of measuring abundance. Furthermore, existing data sets can be used that were originally only intended to deliver protein identifications but no quantitative information. For many purposes, especially related to systems biology, such semi-quantitative information is enough to draw initial conclusions.

3.3.3 Absolute quantification methods

For absolute quantification, the sample is usually spiked with a known concentration of a standard peptide that carries an isotopic label. To achieve precise abundance measurements when comparing two samples, the peptides in the samples are usually also labeled, one with a heavy and one with a light tag. There are several different types of labels and ways of adding them (Shiio and Ruedi Aebersold 2006; Bantscheff et al. 2007).

3.4 Proteomics in *C. elegans*

For the last few years, our lab has undertaken a massive effort to catalogue all measurable proteins in *C. elegans* using different separation techniques. So far, more than half of the predicted proteins in *C. elegans* have already been identified with at least one peptide (see chapter 4). Many gene models were experimentally confirmed for the first time while some could also be corrected.

Even though all life stages have been covered and many different biochemical separation methods have been used, many proteins have still escaped detection. This could be because they are only expressed under certain extreme conditions not encountered in the lab or because the protein abundance is below the detection level of the mass spectrometer. Other possibilities are that some of the predicted genes are in fact incorrectly annotated or that the proteins do not contain any peptides that are measurable in the mass spectrometer.

3.5 Proteomics in other model organisms

The proteomes of several other model organisms have already been published, with a varying degree of coverage. So far, *S. cerevisiae* is the only eukaryote with a virtually complete proteome coverage (de Godoy et al. 2008) (~66%).

Within the QMOP (Quantitative Model Organism Proteomics) initiative, the proteomes of *A. thaliana* (Baerenfaller et al. 2008) (coverage: ~48%), *D. melanogaster* (Brunner et al. 2007) (coverage: ~61%) and *C. elegans* (Schrimpf et al. 2009) have been mapped (see also chapters 4+5, coverage: ~58%).

The complete coverage of the human proteome is the goal of the Human Proteome Organisation (HUPO 2010). The Peptide Atlas (PeptideAtlas 2010) Build for human, which we used for our study (see chapter 5), represents a combination of the efforts of several labs. It covers at the moment about 52% of the predicted proteome.

3.6 Regulation of protein abundance

Many factors contribute to the abundance of proteins as we measure it in the cell. From the initial transcription to protein degradation at the end, there is a complex set of interacting regulators that determine steady-state protein abundance (Davidson 2006; Kevin Chen and Rajewsky 2007). The following is an overview of the most important contributors.

3.6.1 Transcriptional regulation

Unless a gene is silenced, for example by DNA methylation, RNA polymerase can transcribe it to mRNA. This process is regulated via many different mechanisms. Different promoters (or sets of promoters (Heintzman and Ren 2007)) make the binding of RNA polymerase more or less likely. Additionally, specificity factors may change the specificity of this binding.

Gene expression can be reduced by repressors, which bind in the vicinity of the promoter and block transcription. On the other hand, activators may increase the rate of transcription by enhancing the interaction of RNA polymerase and a given promoter.

Generally, the more complex an organism, the more sophisticated the interplay of activators and repressors that finely regulates the rate of transcription (Levine and Tjian 2003).

3.6.2 Post-transcriptional regulation

In eukaryotes, where transcription and translation is spatially separated by the nuclear envelope, mRNAs are extensively modified before translation takes place (Moore 2005).

A cap, consisting of a 7-methylguanosine residue, is added to the 5' end of the transcript right after transcription has started. This cap is essential for recognition by the ribosome and also protects the mRNA from degradation by ribonucleases (Lewis and Izaurralde 1997).

A poly-A tail is added to transcripts at the 3' end, which increases the half-life of the transcript by protecting it from 3' exonucleases. By adjusting the length of the poly-A tail, the cell can determine how long a given mRNA will survive until it is degraded (García-Martínez, Aranda, and Pérez-Ortín 2004).

Splicing modifies the transcript by removing the introns, non-coding stretches of the mRNA. Alternative splicing, i.e. alternative in- or exclusion of coding stretches (exons), allows a single gene to code for many different proteins (Matlin, Clark, and Smith 2005). Recent studies have shown that about 95% of genes with more than one exon undergo alternative splicing (Pan et al. 2008).

A completely new type of regulatory mechanism has become prominent in recent years: microRNAs (miRNAs), typically 22-nucleotide-long non-coding RNAs, which are found in all known animal and plant genomes (David P Bartel 2004). The majority of genes seem to be regulated by one or more miRNAs (Friedman et al. 2009).

In animals, miRNAs usually repress translation by binding with their seed region to a matching complementary sequence in their target mRNA

(possibly with some mismatches), mostly in the 3'UTR. Alternatively, miRNAs might also lead to cleavage of the target mRNA (the main mode of action in plants, mostly by perfect match (Reinhart et al. 2002)) or even upregulate translation (Vasudevan, Tong, and Steitz 2007).

3.6.3 Translational regulation

Translational regulation is likely the most important contributor to the observed large differences between transcript and protein levels (Kelen et al. 2009; de Sousa Abreu et al. 2009). The rate of initiation is the limiting factor in efficiency of translation. It involves a number of initiation factors and is strongly enhanced by the 5' cap and the 3' poly-A tail (Hernández 2009). The Kozak consensus sequence (Kozak 1984), a conserved nucleotide motif around the start codon, also influences the efficiency of ribosome recruitment.

Which codons are used in the transcript affects the rate of elongation; this is why the codon adaptation index correlates well with protein abundance (see chapter 5).

The secondary structure of the transcript also has an effect on translational efficiency as it can slow down translation (Pelletier and Sonenberg 1987).

A means of "storing away" transcripts for later use are the P bodies (Parker and Sheth 2007), aggregates of mRNA-binding proteins, which also seem to be involved in transcript degradation and translational repression by miRNAs.

3.7 Comparison of proteomes

Now that several proteomes with a reasonable coverage are available, we went ahead and asked what we can learn from this information when we compare proteomes with each other, not only qualitatively but also quantitatively.

In a first step (chapter 4), we compared the proteomes of *C. elegans* and *D. melanogaster*. We computed protein abundances via spectral counting for all measured proteins in both organisms and, using orthology information from the STRING database (Jensen et al. 2009), calculated Spearman rank correlation coefficients between the two proteomics data sets as well as with two sets of transcriptome data for each organism.

The surprising result was that protein abundances between organisms correlate even better than transcript with protein abundances within one organism. This means that while protein abundances seem to be quite well conserved, this is not true for transcripts. A low correlation of transcript and protein abundances has been observed before (Gygi et al. 1999; Greenbaum et al. 2003; Xing Fu et al. 2009). For biologically relevant results, one should therefore not rely on transcriptomics data but rather measure the actual protein abundances or otherwise expect limited explanatory power.

Furthermore, when correlating protein abundances, there are functional groups of proteins that show an extremely high correlation coefficient, for example 0.93 for "translation", while obviously there are others with a very low correlation. It is important to note here that we were mostly looking at the well conserved, ubiquitously expressed "core proteome".

To expand on these initial findings, we then collected similar publicly available data of three more organisms: *S. cerevisiae*, *A. thaliana* and *H. sapiens*. The advantage of having several organisms is that we can now also look at the variance of protein abundances across large evolutionary distances (see 3.7.3).

As we show in chapter 5, the high correlations for protein abundances also hold for all pair wise combinations with the other organisms.

3.7.1 Core proteome

When we talk about the "eukaryotic core proteome", we refer to the set of 1172 proteins that have orthologs in all species considered and were also contained in all five proteomics data sets. They are likely to be housekeeping genes, i.e. they should be expressed in almost every cell, under most conditions, at relatively high levels. Nevertheless, we observe a dynamic range of abundances of more than four orders of magnitude.

3.7.2 Evolution of protein abundances

Our results suggest that the abundances of the conserved core proteome are remarkably stable, even over hundreds of million years of evolution. On the other hand, as we could already show in the comparison of fly and worm, the transcript levels are far less conserved. This is likely due to a drift phenomenon for transcripts, i.e. the amount of a given mRNA can change slightly due to mutations which are then compensated by less or more efficient translation, shorter or longer transcript half life, miRNA interaction or similar regulatory mechanisms, so that the protein concentration in the cell remains constant. In the long run, this can lead to wide fluctuations on the mRNA level, and therefore low correlations between organisms, while the protein abundances do not change much, resulting in the observed high correlations.

In recent years, the importance of the complex system of posttranscriptional regulation has become increasingly clear (Lu et al. 2007). This is reflected on a systems level in our findings.

3.7.3 Variance of protein expression

The availability of data from five different organisms allowed us to not only compare abundances, but also examine the variance of expression of conserved proteins. This gives interesting insights into the functional restrictions of protein abundance changes. As we show in chapter 5, there

are indeed groups of proteins with a very low protein abundance variance, which hints at tight regulation, while there are others with a very high variance.

This means that organisms can tolerate quite a bit of expression noise for some groups of proteins, whereas for other groups, the protein concentration in the cell must be kept as constant as possible.

3.8 Publicly available data

As more and more proteomics data are being produced in labs around the world, it becomes more and more important to share these data and make them accessible to the research community (Prince et al. 2004). Several public repositories have been established in recent years, each with a slightly different focus. Obviously, these repositories are only as valuable as the data deposited by researchers. Most proteomics-related journals now encourage authors to submit their data to one of the public repositories and it seems this is soon going to be a requirement before a publication is accepted.

In any case, it is very important that the original raw data files are available because only then can the data be reevaluated and, for example, new search algorithms can be used to re-search the data or it can be searched against an improved protein sequence database. Depending on the requirements, more or less strict quality score cut-offs might be preferred for certain projects.

As we show in chapters 4 and 5, the data can also be used for purposes they were not originally intended for, creating many interesting opportunities for computational biologists.

To allow “recycling” of data, proper annotation is crucial. Otherwise, the deposited data become virtually meaningless. Unfortunately, standards for data storage and annotation are only recently emerging and are not yet very widely used (Martens et al. 2007).

The effort and time invested to create large proteomics data sets makes it all the more essential to share them with the research community at large. Just as biology has enormously benefitted from the free availability of sequenced genomes, the wealth of information contained in these proteomics data sets can only be fully taken advantage of when they are easily accessible to everyone.

3.8.1 Databases and repositories

To ensure that published data does not disappear and can be found without difficulty by interested researchers, several repositories and databases have been set up lately which specialize in proteomics data.

The most important proteomics repositories are the following:

- PeptideAtlas (www.peptideatlas.org)
- PRIDE (www.ebi.ac.uk/pride/)
- Tranche (proteomecommons.org/tranche/)
- Global Proteome Machine (www.thegpm.org)

Each one of them has a certain target audience and a different way of storing and representing the data. Most of these repositories also exchange data with each other (Mead, Shadforth, and Bessant 2007).

3.8.2 Mapping and format problems

A big problem when using data created by somebody else is the mapping of identifiers. Within a model organism community, there is usually at least some consensus which identifiers are the “standard”, like the wormbase CDS for *C. elegans* genes. But even then, it is often not straightforward to map protein names to genes, as annotations change with new releases and some identifiers are renamed or become obsolete.

In the case of human data, different databases, like Ensembl, Swissprot, etc, use a wide variety of different identifiers. As each of these databases has

its own philosophy regarding updating, adding and removing identifiers, there is never a one-to-one mapping.

The different splice variants of proteins represent yet another level of complexity. As the identified peptides usually map to more than one splice variant of a protein, for the time being it is most convenient to not consider different splice variants separately but simply map all peptides to the locus name. We have chosen this approach in our two publications.

In any mapping process, some information is invariably lost. Sometimes, the only solution is to re-search the raw data against a recent release of the relevant protein sequences. This is in turn only possible if the original raw data is still available.

Proteomics data downloaded from any of the repositories usually comes in a proprietary XML format which then has to be parsed to extract the needed information, for example peptide counts. There is often very little documentation for these formats available, so one has to figure out the structure and meaning of the XML by trial and error. This will hopefully soon change as the HUPO proteomics standards initiative (HUPO PSI 2010) is working on XML formats for the representation of proteomics data on all levels (raw, searched as well as analyzed data) that should become the standard file format supported by all repositories (Martens et al. 2007).

3.9 References

Baerenfaller, Katja, Jonas Grossmann, Monica A Grobei, Roger Hull, Matthias Hirsch-Hoffmann, Shaul Yalovsky, Philip Zimmermann, Ueli Grossniklaus, Wilhelm Gruissem, and Sacha Baginsky. 2008. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science (New York, N.Y.)* 320, no. 5878 (May 16): 938-941. doi:10.1126/science.1157956.

Bandeira, Nuno, Jesper V Olsen, Jesper V Mann, Matthias Mann, and Pavel A Pevzner. 2008. Multi-spectra peptide sequencing and its applications to multistage mass spectrometry. *Bioinformatics (Oxford, England)* 24, no. 13 (July 1): i416-423.

Bantscheff, Marcus, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. 2007. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry* 389, no. 4 (October 1): 1017-1031. doi:10.1007/s00216-007-1486-6.

Bartel, David P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, no. 2 (January 23): 281-297.

Brunner, Erich, Christian H Ahrens, Sonali Mohanty, Hansruedi Baetschmann, Sandra Loevenich, Frank Potthast, Eric W Deutsch, et al. 2007. A high-quality catalog of the Drosophila melanogaster proteome. *Nature Biotechnology* 25, no. 5 (May): 576-583. doi:10.1038/nbt1300.

Chen, Kevin, and Nikolaus Rajewsky. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews. Genetics* 8, no. 2 (February): 93-103. doi:10.1038/nrg1990.

Davidson, Eric. 2006. *The regulatory genome : gene regulatory networks in development and evolution*. Amsterdam [Netherlands] ;Boston [Ma.]: Academic Press.

Deutsch, Eric W., Henry Lam, and Ruedi Aebersold. 2008. Data analysis and

bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol. Genomics* 33, no. 1 (October 8): 18-25. doi:10.1152/physiolgenomics.00298.2007.

Dudley, Nathaniel R., and Bob Goldstein. 2005. RNA Interference in Caenorhabditis elegans. In *RNA Silencing*, 29-38. http://dx.doi.org/10.1385/1-59259-935-4:029.

EBI. 2010. EBI Sequenced Genomes Pages - Eukaryota. http://www.ebi.ac.uk/genomes/eukaryota.html.

Friedman, Robin C, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* 19, no. 1 (January): 92-105. doi:10.1101/gr.082701.108.

Fu, Xing, Ning Fu, Song Guo, Zheng Yan, Ying Xu, Hao Hu, Corinna Menzel, et al. 2009. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* 10: 161. doi:10.1186/1471-2164-10-161.

García-Martínez, José, Agustín Aranda, and José E Pérez-Ortín. 2004. Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. *Molecular Cell* 15, no. 2 (July 23): 303-313. doi:10.1016/j.molcel.2004.06.004.

Gevaert, K, and J Vandekerckhove. 2000. Protein identification methods in proteomics. *Electrophoresis* 21, no. 6 (April): 1145-1154. doi:10.1002/(SICI)1522-2683(20000401)21:6<1145::AID-ELPS1145>3.0.CO;2-Z.

de Godoy, Lyris M F, Jesper V Olsen, Jürgen Cox, Michael L Nielsen, Nina C Hubner, Florian Fröhlich, Tobias C Walther, and Matthias Mann. 2008. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 455, no. 7217 (October 30): 1251-1254. doi:10.1038/nature07341.

Greenbaum, D., C. Colangelo, K. Williams, and M. Gerstein. 2003. Comparing

protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* 4, no. 9: 117.

Gygi, S P, Y Rochon, B R Franza, and R Aebersold. 1999. Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology* 19, no. 3 (March): 1720-1730.

Heintzman, N D, and B Ren. 2007. The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome. *Cellular and Molecular Life Sciences: CMLS* 64, no. 4 (February): 386-400. doi:10.1007/s00018-006-6295-0.

Hernández, Greco. 2009. On the origin of the cap-dependent initiation of translation in eukaryotes. *Trends in Biochemical Sciences* 34, no. 4 (April): 166-175. doi:10.1016/j.tibs.2009.02.001.

Hodgkin, Jonathan, H. Robert Horvitz, Barbara R. Jasny, and Judith Kimble. 1998. C. elegans: Sequence to Biology. *Science* 282, no. 5396 (December 11): 2011. doi:10.1126/science.282.5396.2011.

HUPO. 2010. HUPO - Home. http://hupo.org/.

HUPO PSI. 2010. Hupo - Research Projects - Proteomics Standards Initiative. http://hupo.org/research/psi/.

Jensen, Lars J, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, et al. 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research* 37, no. Database issue (January): D412-416. doi:10.1093/nar/gkn760.

Kelen, Katrien Van Der, Rudi Beyaert, Dirk Inzé, and Lieven De Veylder. 2009. Translational control of eukaryotic gene expression. *Critical Reviews in Biochemistry and Molecular Biology* 44, no. 4 (8): 143-168. doi:10.1080/10409230902882090.

Kinchen, Jason M, and Michael O Hengartner. 2005. Tales of cannibalism, suicide, and murder: Programmed cell death in C. elegans. *Current*

Topics in Developmental Biology 65: 1-45. doi:10.1016/S0070-2153(04)65001-0.

Kozak, M. 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Research* 12, no. 2 (January 25): 857-872.

Lange, Vinzenz, Paola Picotti, Bruno Domon, and Ruedi Aebersold. 2008. Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular Systems Biology* 4: 222-222. doi:10.1038/msb.2008.61.

Lettre, Guillaume, and Michael O Hengartner. 2006. Developmental apoptosis in C. elegans: a complex CEDnario. *Nature Reviews. Molecular Cell Biology* 7, no. 2 (February): 97-108. doi:10.1038/nrm1836.

Levine, Michael, and Robert Tjian. 2003. Transcription regulation and animal diversity. *Nature* 424, no. 6945 (July 10): 147-151. doi:10.1038/nature01763.

Lewis, J D, and E Izaurralde. 1997. The role of the cap structure in RNA processing and nuclear export. *European Journal of Biochemistry / FEBS* 247, no. 2 (July 15): 461-469.

Lu, Peng, Christine Vogel, Rong Wang, Xin Yao, and Edward M Marcotte. 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotech* 25, no. 1 (February): 117-124. doi:10.1038/nbt1270.

Marcotte, Edward M. 2007. How do shotgun proteomics algorithms identify proteins? *Nat Biotech* 25, no. 7 (July): 755-757. doi:10.1038/nbt0707-755.

Martens, Lennart, Sandra Orchard, Rolf Apweiler, and Henning Hermjakob. 2007. Human Proteome Organization Proteomics Standards Initiative: Data Standardization, a View on Developments and Policy. *Mol Cell Proteomics* 6, no. 9 (September 1): 1666-1667.

Matlin, Arianne J., Francis Clark, and Christopher W. J. Smith. 2005.

Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology* 6, no. 5 (5): 386-398. doi:10.1038/nrm1645.

Mead, Jennifer A., Ian P. Shadforth, and Conrad Bessant. 2007. Public proteomic MS repositories and pipelines: available tools and biological applications. *PROTEOMICS* 7, no. 16: 2769-2786. doi:10.1002/pmic.200700152.

Moore, Melissa J. 2005. From Birth to Death: The Complex Lives of Eukaryotic mRNAs. *Science* 309, no. 5740 (September 2): 1514-1518. doi:10.1126/science.1111443.

Motoyama, Akira, and John R. Yates. 2008. Multidimensional LC Separations in Shotgun Proteomics. *Analytical Chemistry* 80, no. 19 (October 1): 7187-7193. doi:10.1021/ac8013669.

Pan, Qun, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40, no. 12 (December): 1413-1415. doi:10.1038/ng.259.

Parker, Roy, and Ujwal Sheth. 2007. P Bodies and the Control of mRNA Translation and Degradation. *Molecular Cell* 25, no. 5 (March 9): 635-646. doi:10.1016/j.molcel.2007.02.011.

Pelletier, J, and N Sonenberg. 1987. The involvement of mRNA secondary structure in protein synthesis. *Biochemistry and Cell Biology = Biochimie Et Biologie Cellulaire* 65, no. 6 (June): 576-581.

PeptideAtlas. 2010. PeptideAtlas. http://www.peptideatlas.org/.

Prakash, Amol, Daniela M Tomazela, Barbara Frewen, Brendan Maclean, Gennifer Merrihew, Scott Peterman, and Michael J Maccoss. 2009. Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. *Journal of Proteome Research* 8, no. 6 (June): 2733-2739.

doi:10.1021/pr801028b.

Prince, John T, Mark W Carlson, Rong Wang, Peng Lu, and Edward M Marcotte. 2004. The need for a public proteomics repository. *Nature Biotechnology* 22, no. 4 (April): 471-472.

Reinhart, Brenda J., Earl G. Weinstein, Matthew W. Rhoades, Bonnie Bartel, and David P. Bartel. 2002. MicroRNAs in plants. *Genes & Development* 16, no. 13 (July 1): 1616-1626. doi:10.1101/gad.1004402.

Reiter, Lukas, Manfred Claassen, Sabine P. Schrimpf, Marko Jovanovic, Alexander Schmidt, Joachim M. Buhmann, Michael O. Hengartner, and Ruedi Aebersold. 2009. Protein identification false discovery rates for very large proteomics datasets generated by tandem mass spectrometry. *Mol Cell Proteomics* (July 16): M900317-MCP200. doi:10.1074/mcp.M900317-MCP200.

Schrimpf, Sabine P, Manuel Weiss, Lukas Reiter, Christian H Ahrens, Marko Jovanovic, Johan Malmström, Erich Brunner, et al. 2009. Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biology* 7, no. 3 (March 3): e48. doi:10.1371/journal.pbio.1000048.

Shiio, Yuzuru, and Ruedi Aebersold. 2006. Quantitative proteome analysis using isotope-coded affinity tags and mass spectrometry. *Nature Protocols* 1, no. 1: 139-145. doi:10.1038/nprot.2006.22.

de Sousa Abreu, Raquel, Luiz O Penalva, Edward M Marcotte, and Christine Vogel. 2009. Global signatures of protein and mRNA expression levels. *Molecular bioSystems* 5, no. 12 (December): 1512-1526. doi:10.1039/b908315d.

Steen, Hanno, and Matthias Mann. 2004. The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews. Molecular Cell Biology* 5, no. 9 (September): 699-711. doi:10.1038/nrm1468.

Vasudevan, Shobha, Yingchun Tong, and Joan A Steitz. 2007. Switching from

repression to activation: microRNAs can up-regulate translation. *Science (New York, N.Y.)* 318, no. 5858 (December 21): 1931-1934. doi:10.1126/science.1149460.

Wood, William B. 1988. *The Nematode Caenorhabditis Elegans*. CSHL Press, June.

4 Comparative Functional Analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* Proteomes

4.1 Preface

For this publication, I contributed most of the computational analysis of the data, especially the prediction and distribution of transmembrane domains, the assignment of Gene Ontology terms and their mapping to higher-level GOslim terms, as well as an analysis of the distribution of these terms.

Furthermore, I developed and implemented an algorithm to estimate protein abundances based on the idea of spectral counting, did extensive qualitative performance testing and computed abundances for all *C. elegans* and *D. melanogaster* proteins.

I also performed all the correlation analyses and made the corresponding figures.

Comparative Functional Analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* Proteomes

Sabine P. Schrimpf[1,2*], Manuel Weiss[1,2,3], Lukas Reiter[1,2,3,4], Christian H. Ahrens[2,5], Marko Jovanovic[1,2,3], Johan Malmström[4], Erich Brunner[2], Sonali Mohanty[2,4], Martin J. Lercher[6], Peter E. Hunziker[5], Ruedi Aebersold[4,7], Christian von Mering[1,2,8*], Michael O. Hengartner[1,2*]

1 Institute of Molecular Biology, University of Zurich, Zurich, Switzerland, **2** Center for Model Organism Proteomes, University of Zurich, Zurich, Switzerland, **3** PhD Program in Molecular Life Sciences, University of Zurich, Zurich, Switzerland, **4** Institute of Molecular Systems Biology, Swiss Federal Institute of Technology Zurich, Zurich, Switzerland, **5** Functional Genomics Center, University of Zurich and Swiss Federal Institute of Technology Zurich, Zurich, Switzerland, **6** Institute of Informatics, University of Düsseldorf, Düsseldorf, Germany, **7** Institute for Systems Biology, Seattle, Washington, United States of America, **8** Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland

The nematode *Caenorhabditis elegans* is a popular model system in genetics, not least because a majority of human disease genes are conserved in *C. elegans*. To generate a comprehensive inventory of its expressed proteome, we performed extensive shotgun proteomics and identified more than half of all predicted *C. elegans* proteins. This allowed us to confirm and extend genome annotations, characterize the role of operons in *C. elegans*, and semiquantitatively infer abundance levels for thousands of proteins. Furthermore, for the first time to our knowledge, we were able to compare two animal proteomes (*C. elegans* and *Drosophila melanogaster*). We found that the abundances of orthologous proteins in metazoans correlate remarkably well, better than protein abundance versus transcript abundance within each organism or transcript abundances across organisms; this suggests that changes in transcript abundance may have been partially offset during evolution by opposing changes in protein abundance.

Citation: Schrimpf SP, Weiss M, Reiter L, Ahrens CH, Jovanovic M, et al. (2009) Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. PLoS Biol 7(3): e1000048. doi:10.1371/journal.pbio.1000048

Introduction

The rapid lifecycle, small size, reproducible development, and ease of cultivation in the laboratory have made *Caenorhabditis elegans* an important experimental system for biological studies. Numerous human disease-related genes (e.g., related to cancer or neurological diseases) have orthologs in the worm [1]. Sequencing and annotation of its genome has revealed more than 19,000 genes [2] coding for more than 22,000 proteins, including splice variants. Extensive systematic studies of gene function have been performed. However, to completely understand complex biological processes such as development, aging, or disease, the analysis of the proteome—i.e., the entire set of the expressed proteins—is becoming increasingly important. Knowledge of the complete sequence of a genome is a necessary prerequisite for proteomics, but the DNA sequence itself does not reveal which proteins are actually expressed when, where, and to what level. Furthermore, in contrast to the genome, the proteome is changing under different biological conditions. Although for many years, transcriptome data (i.e., the collection of transcribed mRNAs) has been used to approximate the proteome, a number of studies have demonstrated that the correlation between mRNA and protein abundance is surprisingly low [3–5] because of posttranscriptional regulation and variable protein half-lives. The analysis of the proteome is therefore a key method to provide systems-level information about protein function in time and space, and to obtain a concise view of biological processes. In the case of *C. elegans*, previous analyses of the proteome were either limited in scope and coverage [6,7], or largely focused on improving genome annotation [8], with the biggest *C. elegans* proteome dataset published so far encompassing 6,779 proteins [8].

To generate a comprehensive, deeply sampled *C. elegans* proteome database that can be used for quantitative proteome analysis, we applied subcellular and biochemical fractionation methods to the worm proteins, performed tryptic digests, separated the resulting peptides using a variety of techniques, and identified the peptides by mass spectrometry (MS). This resulted in a unique global view on the expression status of the *C. elegans* proteome. We identified a number of protein features and functions that are underrepresented in the expressed proteome, likely representing specialized functional systems expressed only in a small subset of cells and/or developmental stages. We demonstrate the importance of proteomics data towards improved genome annotation. Finally, we compared the proteome data with similar data from the fruit fly *Drosophila melanogaster*. The

Academic Editor: Jonathan S. Weissman, University of California San Francisco and Howard Hughes Medical Institute, United States of America

Received June 11, 2008; **Accepted** January 13, 2009; **Published** March 3, 2009

Abbreviations: FDR, false discovery rate; GO, Gene Ontology; MS, mass spectrometry; MS/MS, tandem mass spectrometry; pI, isoelectric point; SAGE, serial analysis of gene expression

* To whom correspondence should be addressed. E-mail: sabine.schrimpf@molbio.uzh.ch (SPS); mering@molbio.uzh.ch (CvM); michael.hengartner@molbio.uzh.ch (MOH)

Author Summary

Proteins are the active players that execute the genetic program of a cell, and their levels and interactions are precisely controlled. Routinely monitoring thousands of proteins is difficult, as they can be present at vastly different abundances, come with various sizes, shapes, and charge, and have a more complex alphabet of twenty "letters," in contrast to the four letters of the genome itself. Here, we used mass spectrometry to extensively characterize the proteins of a popular model organism, the nematode *Caenorhabditis elegans*. Together with previous data from the fruit fly *Drosophila melanogaster*, this allows us to compare the protein levels of two animals on a global scale. Surprisingly, we find that individual protein abundance is highly conserved between the two species. So, although worms and flies look very different, they need similar amounts of each conserved, orthologous protein. Because many *C. elegans* and *D. melanogaster* proteins also have counterparts in humans, our results suggest that similar rules may apply to our own proteins.

latter comparison provided—for the first time to our knowledge—an overview of the expressed "core animal proteome," which should arguably become the initial focus for monitoring the basic metazoan cellular machinery in the future.

Results

Protein Identifications

To identify *C. elegans* proteins, we collected worms at various developmental stages and homogenized whole animals and eggs to isolate the proteins. Their tryptic peptides were separated using strong cation exchange chromatography (SCX), in several cases after labeling them with isotope-coded affinity tags (ICAT) [9] to reduce sample complexity, or by isoelectric focusing (applying free-flow electrophoresis and immobilized pH gradient strips). The peptides were finally identified using microcapillary liquid chromatography–electrospray ionization–tandem MS (μLC-ESI-MS/MS). With this extensive shotgun proteomics approach, we identified 10,977 different proteins, including splice variants, via 84,962 nonredundant peptide identifications (Table S1; 759,320 peptide identifications were obtained in total). We identified 10,631 gene loci, corresponding to 54% of the gene loci in WormBase (WS140: 19,735 loci). Of these, 7,476 loci (38%) were detected via several distinct peptides, 580 (3%) were detected via the same peptide more than once, and 2,575 (13%) were detected only via a single peptide identification (Figure 1). When considering individual annotated exons (irrespective of their various splicing contexts), our peptide data covered 28.2% of the 129,047 exons contained in WormBase.

Protein identification from MS peptide spectra is prone to false-positive assignments, and we employed strict search cutoffs using PeptideProphet (see Materials and Methods). To independently estimate our false discovery rate (FDR), in particular for identifications based on a single peptide spectrum ("single hits"), we first took advantage of one of our experiments that used isoelectric focusing to fractionate peptides. In each peptide fraction, true-positive identifications should scatter around a narrow range of isoelectric points (pIs), whereas false-positive identifications should follow the background distribution in the database. This analysis, using computational predictions of pIs to check all peptides, yielded an estimated FDR of 35% for single hits in this particular experiment. Independently, a newly developed model based on a robust decoy search strategy yielded an upper limit for the FDR of single-hit identifications at around 63% for all combined experiments (L. Reiter, M. Claassen, S. P. Schrimpf, J. M. Buhmann, M. O. Hengartner, et al., unpublished data). By the latter method, multi-hit identifications were found to be much more reliable, resulting in an FDR of 7% in our study. Since almost half of all single-hit identifications do represent bona fide protein identifications, we chose to include single-hit identifications in our subsequent analyses. A separate analysis focusing on just these proteins alone showed that they often belonged to groups that were underrepresented in the complete dataset and are therefore presumably of low abundance in *C. elegans* (short, uncharacterized proteins and in particular those with seven

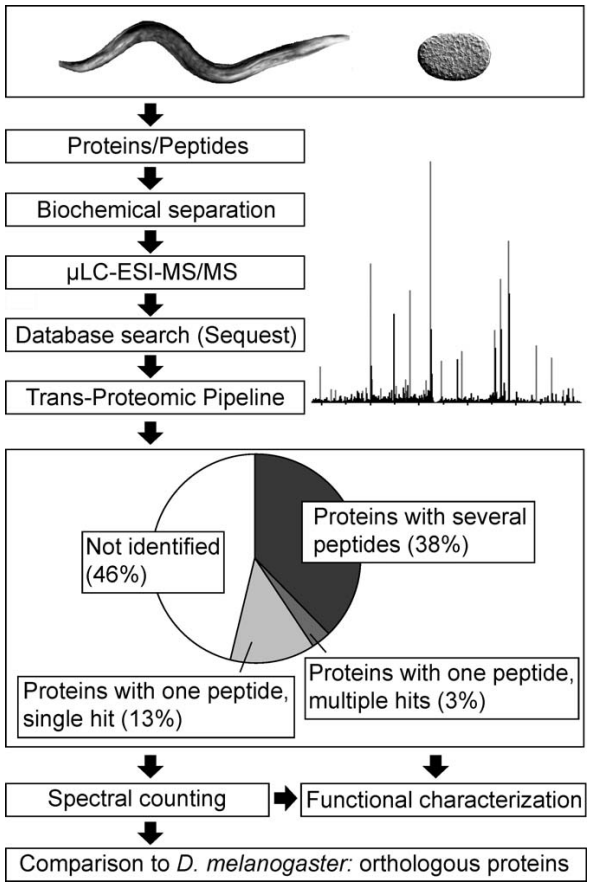


Figure 1. Workflow of the *C. elegans* Proteome Analysis

Proteins and peptides were isolated from whole worm or egg homogenates, and separated biochemically. Peptides were identified by μLC-ESI-MS/MS and database searches, and validated using the Trans-Proteomic Pipeline [62]. We detected peptides for 10,631 different gene loci, which corresponds to 54% of the predicted gene loci in WormBase WS140 (19,735 gene loci). For 7,476 gene loci, more than one peptide was identified; for 580 gene loci, a single peptide was identified independently multiple times; for 2,575 gene loci, a single peptide was identified; and 9,104 gene loci were not covered at all.

doi:10.1371/journal.pbio.1000048.g001

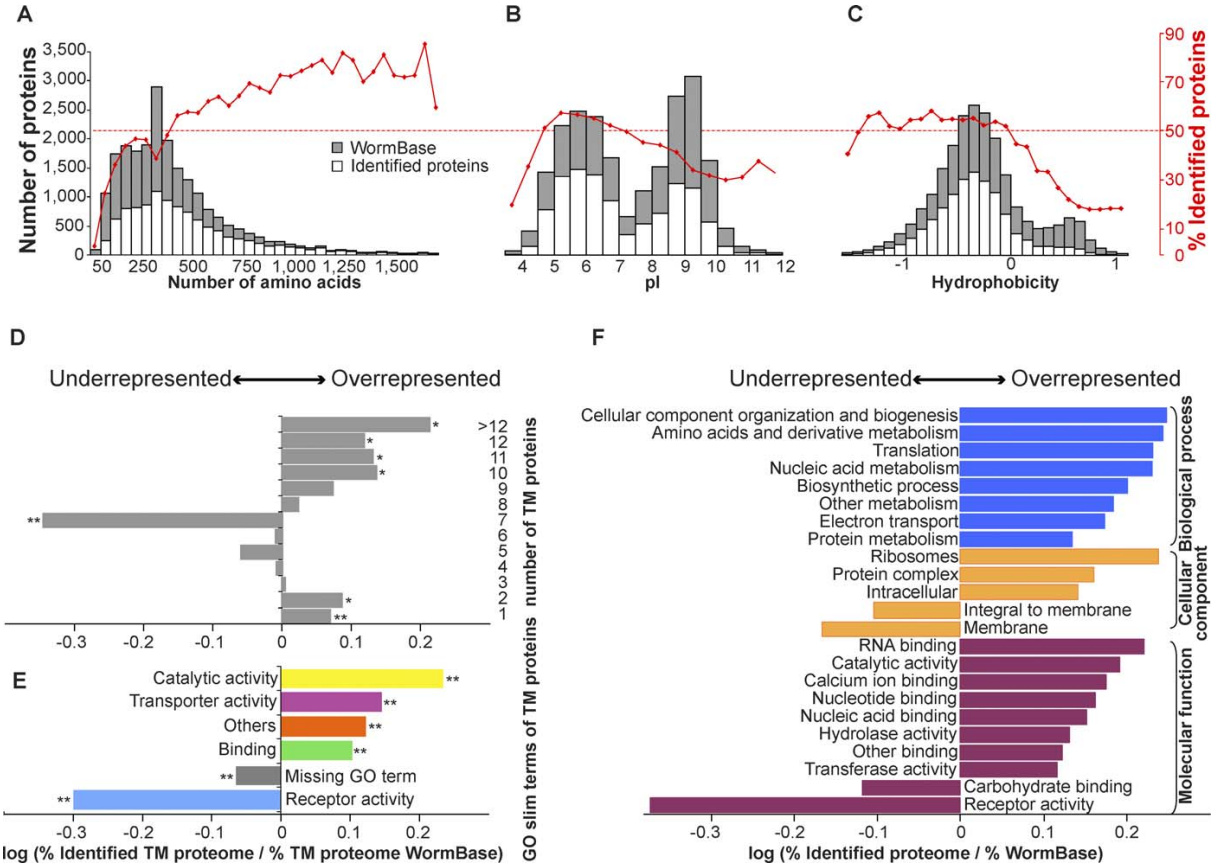


Figure 2. Classification of Detected Proteins

(A–C) A bias analysis of the 10,977 identified proteins (including splice variants) in comparison to the 22,269 predicted proteins in WormBase (WS140) was performed for the parameters (A) length, (B) isoelectric point (pI), and (C) hydrophobicity. Red lines indicate the percentages of identified proteins in comparison to all *C. elegans* proteins in each bin. A value below 49% indicates fewer detections than expected; a value above 49% indicates more detections than expected.

(D and E) Over- and underrepresentations of transmembrane (TM) proteins (D) and their functional classes (E) in our experimental dataset. Statistically significant categories are labeled with asterisks: *p*-values better than 0.05 are indicated by a single asterisk (*); *p*-values better than 1E–4 are indicated by double asterisks (**). The proportion of proteins with transmembrane helices was 36.5% in WormBase, and 30.5% in our proteome dataset.

(F) The global functional GO slim analysis for all proteins showed statistically significant over- or underrepresentations in the categories "biological process," "cellular component," and "molecular function." We used abbreviated terms for three categories (GO:0006139, GO:0008152, and GO:0005488).

doi:10.1371/journal.pbio.1000048.g002

transmembrane domains; also see below). This means that they do represent valuable information about which proteins are expressed at low levels in *C. elegans*. It should also be stressed that all conclusions reported below remained valid when single-hit identifications were excluded.

To assess whether proteins from sources other than *C. elegans* were present in our preparations, we focused on the bacteria on which the worms were feeding (*Escherichia coli*). We tested a single, representative experiment, encompassing 67 MS/MS analyses by searching the spectra against a combined *C. elegans* and *E. coli* database. A total of 1.3% of the protein identifications mapped to *E. coli*, among them 14 hits mapping to both organisms. However, for each of these 14 proteins, there was at least one additional *C. elegans* peptide identified, confirming that these overlapping detections did not influence the *C. elegans* results.

Proteins Seen and Not Seen: Features and Functions

In order to characterize *C. elegans* proteins that were *not* detected, and that are therefore most likely expressed at particularly low levels, or in specialized cells or developmental stages only, we classified the entire predicted *C. elegans* proteome with respect to several aspects (length, pI, hydrophobicity, transmembrane topology, and functional annotation). This should reveal the nature of underrepresented proteins (with potentially more peripheral, or even worm-specific functions), and separate them from abundant proteins involved in basic cellular processes such as growth, metabolism, and information processing. It should also reveal potential technical limitations (proteins/peptides difficult to detect using our procedure), which is important to assess for future systematic uses of MS.

Our bias analyses revealed an underrepresentation of proteins shorter than 400 amino acids (Figure 2A) and of

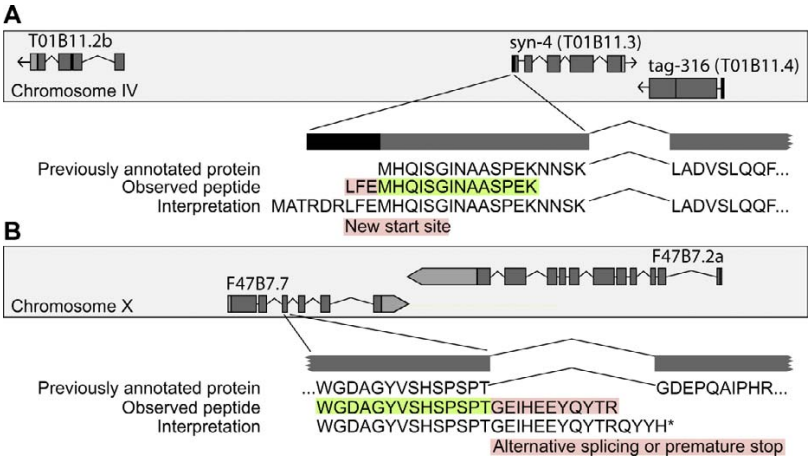


Figure 3. Improved Genome Annotation via Novel Peptide Identifications

Examples of novel peptides obtained from genomic searches against a six-frame translation of the *C. elegans* genome, and the region where they match to the genome.

(A) The novel peptide sequence LFEMHQISGINAASPEK suggests an alternative translational start site for the protein SYN-4 (T01B11.3). The sequence predicted to code for this peptide extends upstream of the annotated translational start site. An alternative start codon can be found further upstream in the same reading frame.

(B) A peptide points at a novel splice variant that was identified for the gene F47B7.7. The peptide WGDAGYVSHSPSPTGEIHEEYQYTR extends an existing annotated exon into the downstream intron, resulting either in the selection of an alternative 5′ splice site downstream of the peptide, or in intron retention, which would result in an early translation stop (shown).

doi:10.1371/journal.pbio.1000048.g003

proteins with basic pIs (Figure 2B). A similar bias has previously been observed for *D. melanogaster* [10]. The underrepresentation of basic proteins was partly to be expected, due to our isoelectric focusing experiments, which centered on the pH range 3–7. The underrepresentation of short (low molecular weight) proteins might be caused by a generally higher prevalence of spurious gene predictions among short genes, and also by a lower probability of detecting one of the few tryptic peptides generated by short proteins. We observed a bimodal distribution of hydrophobicity values within the annotated set of all *C. elegans* proteins, and a strong underrepresentation of proteins in the second, high hydrophobicity peak in our dataset (Figure 2C). This second peak consists mostly of multipass transmembrane proteins (~64% of these proteins have seven or more predicted transmembrane domains). To better understand how membrane association relates to protein abundance and detectability, we globally characterized WormBase proteins with respect to their content of transmembrane segments, using Phobius [11]. Overall, we found a notable underrepresentation of transmembrane proteins in our proteomics data, and decided to subdivide these proteins further according to the number of transmembrane sections and annotated functions as shown for other species [12,13] (Figure 2D and 2E). Remarkably, we found that the strongest underrepresentation is observed for proteins with seven transmembrane regions, in particular those annotated with the function "receptor activity." This may point to a biological (rather than technical) explanation for the relative paucity of transmembrane proteins in our data: Seven-transmembrane chemosensory receptors are widespread in the *C. elegans* genome, but many of these are known to be expressed only in a small number of neurons each [14–16]. Because we assessed whole animals, those proteins might be too rare to be successfully detected. This general underrepresentation in our proteome data suggests similar sensory functions for other transmembrane proteins of hitherto unknown function that we also found to be of too low abundance to be detected.

Finally, we globally analyzed the functional classifications of all the detected proteins. We observed a clear bias towards proteins with known functions. The same bias was also observed for the *D. melanogaster* proteome [10]. A possible explanation could be that some of the undetected proteins with unknown functions are actually erroneous gene predictions or pseudogenes. It could also be a testament to the biases of previous studies: abundant proteins are easier to work with biochemically, and may therefore have obtained a functional annotation more easily. In total, our proteomics approach identified proteins belonging to 125 out of the 127 Gene Ontology (GO) slim categories defined for WormBase. The global GO slim analyses confirmed the underrepresentation of proteins with receptor activity mentioned above, and of "membrane" or "integral to membrane" proteins in general (Figure 2F).

Improving Genome Annotation

Large-scale proteome analyses (such as ours) represent an important cornerstone for an improved genome annotation. In WormBase (WS160), 4,987 gene loci were still listed with the gene status "predicted" only, i.e., without any supporting transcript data (expressed sequence tag [EST], mRNA). We experimentally confirmed the protein expression of 1,062 of these predicted genes (among them, more than 40% via multiple peptide detections). As was the case for the whole proteome, this subset was enriched for proteins with GO slim annotations (45% in our dataset, as compared to 38% expected for this subset in WormBase; *p*-value: 4.65E–08). Apart from these gene confirmations, our *C. elegans* proteomics dataset contains numerous spectra originating from

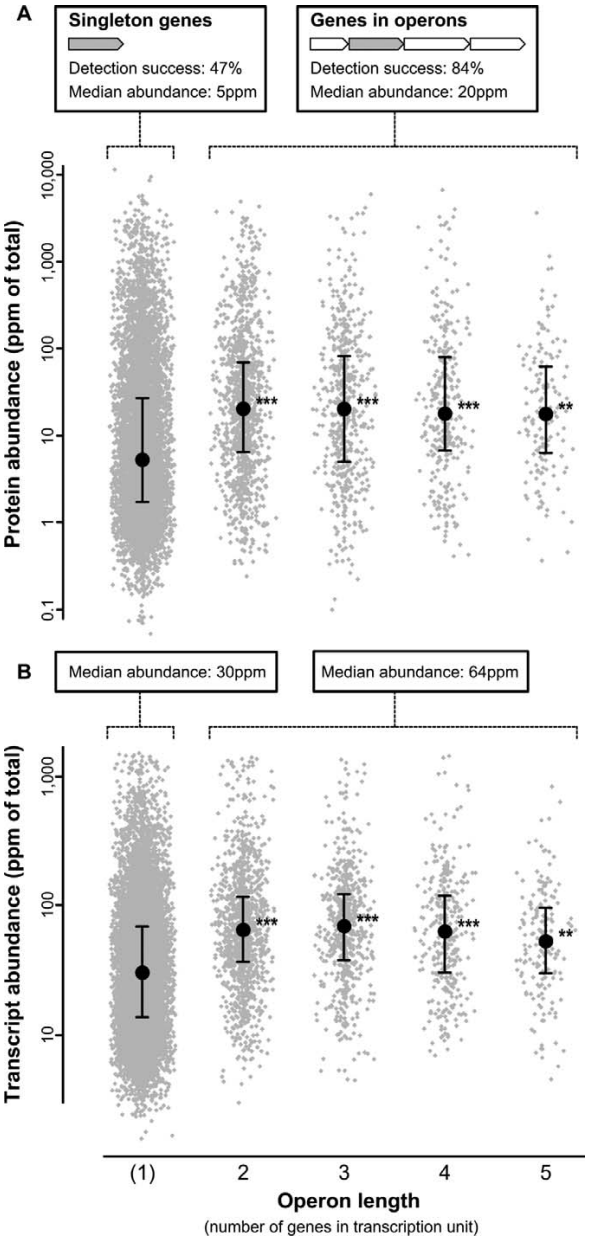


Figure 4. Operon Genes Are More Highly Expressed Than Singleton Genes

(A) Proteins whose genes are organized in operons were identified more frequently (84%) and more abundantly (median expression: 20 ppm) compared to proteins encoded by individually transcribed genes (47%; 5 ppm). *p*-values: double asterisks (**) indicate better than 1E–10; triple asterisks (***) indicate better than 1E–15.

(B) A similar result is obtained when analyzing Affymetrix data instead (albeit with a smaller abundance difference). In both panels, the left-most data column encompasses singleton genes (i.e., not in operons), and the four columns to the right encompass genes in operons of various lengths. Medians are indicated as black dots, and whiskers encompass the range from 25% to 75% of values.

doi:10.1371/journal.pbio.1000048.g004

nonannotated regions in the worm genome. In computationally intensive analyses, we are identifying these by searching our data against six-frame translations of the genome, and filtering the results for high confidence spectra that map to nonannotated regions. For example, from one particular experiment, we identified 78 likely novel peptides. Two of these are illustrated in Figure 3 (the corresponding MS/MS spectra are provided as Figure S1). These data suggest an alternative translational start site for the protein SYN-4 (T01B11.3; Figure 3A): the observed peptide is located upstream of the annotated translational start site, and only partially overlaps with the currently annotated protein sequence. The second example demonstrates a novel splice variant for the gene F47B7.7 (Figure 3B). In this case, we identified a peptide that extends an existing annotated exon downstream, in the correct frame. These and similar analyses, suggesting altered or new gene models, are computationally very intensive and were not yet completed at the time of submission. Furthermore, due to the increased search space when searching proteomics against the genome, extra scrutiny is needed when interpreting each reannotation instance, and additional experimental data should probably be taken into account before fully accepting these gene annotation changes.

Operons

C. elegans and its relatives are unique among characterized metazoans in that a large number of their genes are organized into operons (multicistronic transcription units, containing up to eight genes that are strictly coexpressed [17,18]). Following transcription, the primary transcript is split up through a unique trans-splicing mechanism, and the individual open reading frames are subsequently translated separately into distinct mature proteins. In order to assess the potential influence of operon structure on the regulation and abundance of proteins, we studied the expression status of genes in operons, compared to individually transcribed genes. Although an absolute quantification of protein levels is not possible with our shotgun approach, we performed a semiquantitative analysis based on spectral counting [19–23]. Surprisingly, we observed that proteins encoded by operons are expressed far more strongly than those encoded by individually transcribed genes: we observed 84% of the former, with a median relative abundance of 20 ppm (parts per million of total protein molecules), but only 47% of the latter with a median relative abundance of 5 ppm (Figure 4A). The same tendency was found when analyzing publicly available transcript-abundance data (Figure 4B). This striking observation confirms that operons are preferentially made up of genes that are strongly transcribed, and we now establish that this is reflected also at the protein level: operon proteins, on average, are more than 3-fold more abundant than proteins from single-gene transcripts. Apart from grouping strongly expressed proteins, operons are also expected to facilitate coordinated regulation of their constituent genes. We assessed whether this is the case by searching for operons that were either fully expressed (i.e., all encoded proteins detectable) or silenced (none or very few of the encoded proteins detectable). Indeed, we found significantly more operons of both types than expected by chance, as illustrated for operons of lengths 4 to 6 (Figure S2). In principle, our observations could be stemming from a limited selection of tissues only, for example from the hermaphrodite germline,

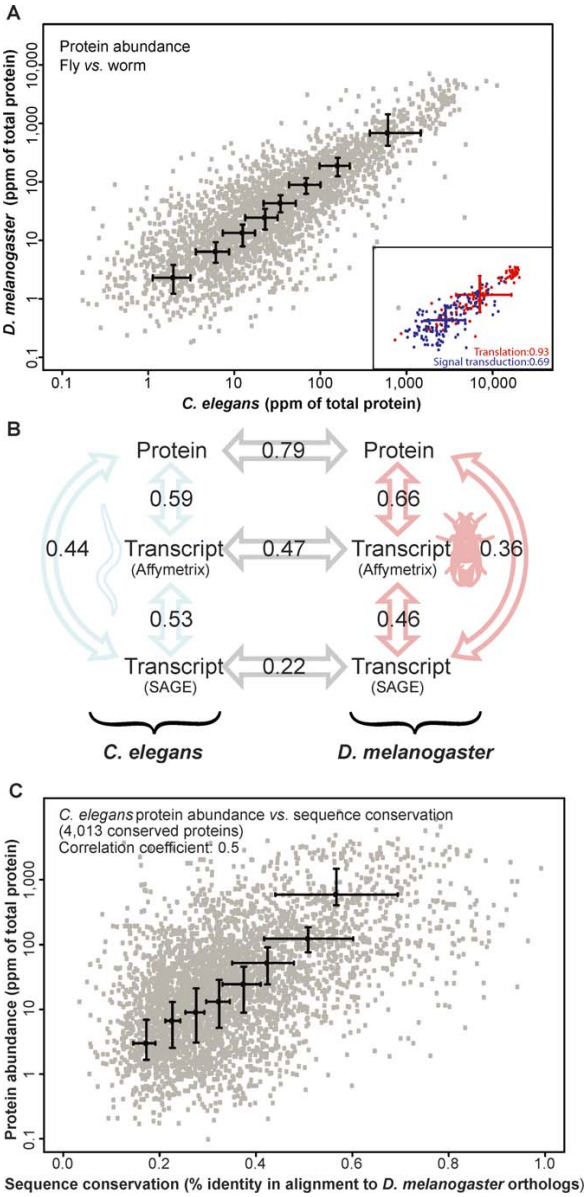


Figure 5. Interspecies Comparative Proteomics of Orthologous Proteins in *C. elegans* and *D. melanogaster*

(A) Protein abundances deduced from spectral counting of 2,695 pairs of orthologs from both species are shown. Medians of equal-sized bins are indicated as crosses; whiskers encompass the range from 25% to 75% of values. The distribution of the orthologs (dots) is indicated in the background. The distribution and correlation coefficients of proteins involved in signal transduction and translation are shown in the inset.
(B) The correlation coefficient of $R_S = 0.79$ between the two species is higher than that of the comparison between protein and transcript abundance within the organisms, based on SAGE or Affymetrix data.
(C) For *C. elegans*, we plotted protein abundance versus sequence conservation (the latter determined by alignment with the *D. melanogaster* orthologs). All correlation coefficients are rank-based with *p*-values better than 2.2E−16.
doi:10.1371/journal.pbio.1000048.g005

where operons are thought to be strongly expressed during oogenesis [24]. However, we observed that operon proteins are more abundant even in dauer and L1-stage larvae, which both should have very little germline material (Figure S3). We further checked whether our observation could be explained by systematic differences in length or transmembrane segments of operon proteins. Although we did observe slight differences in length and transmembrane content—operon proteins are on average 11% longer, and transmembrane proteins are 40% less frequent—these differences were not sufficient to explain the increased abundance of operon proteins (unpublished data). Together, our observations indicate, for the first time, that operons in *C. elegans* ensure the coordinated regulation of highly expressed proteins.

Comparison to the *D. melanogaster* Proteome Dataset

In this study, we had the unique opportunity to compare large-scale proteome datasets from two different animal species, owing to the recent publication of the *D. melanogaster* proteome [10] (http://www.mop.uzh.ch/peptideatlas/; previous work in *D. melanogaster* had mainly focused on protein–protein interactions or subproteomes only [25,26]). We performed spectral counting for both organisms to obtain semiquantitative measurements of protein abundance, and compared these to published mRNA expression data derived from Affymetrix [27,28] and serial analysis of gene expression (SAGE) platforms [27,29]. In the *C. elegans* and *D. melanogaster* proteomes, 2,695 pairs of orthologs were identified for which all three types of data were available. Surprisingly, we observed that orthologs showed a strong correlation in protein abundances across the two organisms, despite more than 600 million years of separate evolution (Spearman rank correlation $R_S = 0.79$; Figure 5A). Notably, this biological correlation at the protein level between the two species is even higher than the within-species correlation between protein and transcript abundances (within *C. elegans*: $R_S = 0.59$ and 0.44 for protein-Affymetrix and protein-SAGE, respectively; within *D. melanogaster*: $R_S = 0.66$ and 0.36, respectively). In contrast to the protein-level correlation, the abundance correlations at the transcript level between the two species were also rather low (Figure 5B). Interestingly, the overall protein-abundance correlations are not equally tight across functional categories: the highest correlation was observed for the functional category "translation" ($R_S = 0.93$) and the lowest for the category "regulation of biological process" ($R_S = 0.65$).

Despite the fact that it is difficult to compare tissues and developmental stages across organisms, our analysis provides a first insight into the evolutionary behavior of animal proteins over long time scales. It is important to point out that for all six data points, several developmental stages and/or tissues had been mixed, but that these were not, of course, always directly equivalent and comparable between the two organisms. However, many of the ancient animal orthologs that we studied here can be expected to be expressed similarly across many cell types and stages, and we thus attempted to capture an organism-wide "average" proteome for both animals. That notwithstanding, we also repeated the analysis for one set of samples that is arguably more directly comparable: mixed staged embryos sampled in both *D. melanogaster* and *C. elegans* at the proteome and at the transcript levels (Figure S4). Here again, we saw that protein abundances correlated far better ($R_S = 0.70$) across organisms than transcript abundances ($R_S = 0.50$).

Another potential complication for our analysis lies in the technique of spectral counting. Individual tryptic peptides are known to ionize and be detected with widely differing efficiencies in mass spectrometers. Although protein conservation between *C. elegans* and *D. melanogaster* is low (~40% sequence identity), a higher-than-expected abundance correlation might still result if equivalent peptides in both organisms were correlated in their suitability for MS. We assessed the extent of this effect by making the spectral counts independent: For any given section in the alignment of two orthologs, only one of the proteins was allowed to generate peptide counts; these sections were alternated across the length of the alignment, effectively reducing the data by half. As expected, this lowered the abundance correlation, but not by much ($R_S = 0.68$). Importantly, the resulting correlation is still much higher than the correlation of transcript abundances across organisms (Figure S5).

Since one of our original interests was to characterize the "core animal proteome," we also analyzed lower-coverage datasets from two additional organisms: *Saccharomyces cerevisiae* [30] and *Mus musculus* ([21]) (for the latter, we additionally included plasma data from PeptideAtlas; http://www.peptideatlas.org/). Comparative proteomics using multiple organisms has recently become popular, for example in bacteria [31], but has not yet been possible for animals. We searched for groups of orthologs that were detected in all four organisms; these would constitute the universally detectable eukaryotic proteome core. We found 847 such proteins, mostly from information-processing and metabolism genes. Conversely, we found 1,287 proteins to be detectable in all three animals, but not in yeast. This latter set might be considered the specific core of multicellular animal proteomes. However, it is clear that neither of these sets is complete, as of yet, mostly due to low coverage in mouse.

Expression Levels of Duplicated Genes

Our protein-abundance estimates from two organisms also allowed us to study in more detail the fate of duplicated genes. Here, of particular interest, are cases in which a gene family has duplicated in one lineage (fly or worm), but not in the other. It is known that long-term retention of duplicated gene copies requires neo- or subfunctionalization [32–34], but it is unclear what consequences this has for overall protein-abundance levels. We found that when averaging over all cases of lineage-specific gene duplications (Figure S6), the abundance of duplicated genes is significantly lower than that of their nonduplicated counterparts in the other lineage. Strikingly, however, when all the duplicated genes of a given gene family are pooled, they tend to add up again to the original abundance of the nonduplicated counterpart (Figure S6).

Discussion

We describe here a comprehensive inventory of *C. elegans* proteins, the functional characterization of this inventory, and the first-ever comparison of two such inventories between two model animals ("comparative proteomics"). Although some subsets of the proteome are more difficult to analyze (e.g., the membrane compartment), we achieved a relatively thorough representation of the genome, where the major exceptions can be explained biologically. For example, the systematic underrepresentation of seven-transmembrane proteins appears to be caused mainly by G protein–coupled receptors. The putative chemoreceptor gene families in *C. elegans* encompass about 7% of its total genome [35], and many are thought to be expressed only in a few neurons each [14–16]. Despite their generally low abundance, we did identify 172 seven-transmembrane receptor proteins, showing that they are, in principle, amenable to high-throughput MS analysis (this is relevant, for example, for screens of putative therapeutic targets).

We also demonstrated that a whole-proteome analysis of a model organism can contribute to an improved genome annotation. First, we experimentally confirmed the expression of 1,062 predicted genes for which no transcript data were available, but for which our proteome data allowed the extraction of a first rough expression pattern. Second, we identified novel peptides from spectra that could not be matched to annotated gene models, suggesting a way to more precisely map open reading frames and splice isoforms to the genome.

With respect to genome organization, we found that, in *C. elegans*, genes in operons are far more consistently and more strongly expressed than individually transcribed genes. In principle, this observation could be an artifact of genome annotation—if a disproportionally large number of annotated nonoperon genes were misannotations that are biologically meaningless. This is highly unlikely, however, since more than 6,000 such misannotations would be needed to reconcile the observed differences. Instead, it is likely that operons in *C. elegans* indeed serve to group strongly expressed genes into coregulated transcription units. Another question that arises is whether these genes were highly expressed even *before* they were grouped into operons, which would hint at a possible selective advantage for the grouping (e.g., to enable more efficient, more reliable, or more uniform transcription of genes whose products are in high demand). This is difficult to address conclusively, but our comparison to *D. melanogaster* provides some information: we observe that orthologs of operon genes are more strongly expressed even in the fly (Figure S7), where they are not arranged in operons nor are even neighbors on the genome. If one assumes that the operons in *C. elegans* are the derived state, then the corresponding genes were indeed already strongly expressed before they formed operons.

The comparison of our data to the *D. melanogaster* proteome also sheds some light on an important evolutionary puzzle, namely the surprisingly low correlation between mRNA expression levels of orthologous genes across animal species [36,37], despite evidence for strong stabilizing selection against expression changes in experimental evolution [38]. We found that the abundances of orthologous proteins from worm and fly correlate well ($R_S = 0.79$), far better than the corresponding abundances of mRNA transcripts ($R_S < 0.50$; Figure 5B).

There are several possible explanations for this finding: First, sweeping changes within the transcriptional machineries in one or both organisms could have resulted in global differences in transcript abundance, whereas selection would have kept protein abundances at least partially stable. One candidate for such a mechanistic change could be, for example, the unique trans-splicing mechanism of nematodes. A second possible explanation might be that posttranslational regulation may have changed systematically, for example due to differences in developmental strategies, physiology, or life styles of the two

animals. Here, possibly relevant changes include the fixed cell lineage of nematodes, differences in reproductive strategies, increased endurance in nematodes (dauer stage), or the constraints imposed on *D. melanogaster* because of its need for metamorphosis and its higher motility (flight).

However, in our view, the most parsimonious explanation might be that many changes in the transcriptome might be neutral, or at least nearly neutral [36]. Ultimately, it is the protein levels that are under selection. Protein levels are not only determined by mRNA abundance, but are equally affected by translation efficiencies, protein half-lives, and other factors. Genetic mutations resulting in small changes on any of these levels might persist for some time in a population, as long as their fitness effects are small (around $1/[2N_e]$ or less). This might be sufficient time to allow for compensatory mutations either in the same gene or elsewhere in the genome, which would reconstitute optimal protein abundance through action on the same or another factor that influences protein abundance. Thus, changes in mRNA expression could be offset by opposite changes in translation rate or protein half-life, and vice versa. Over evolutionary time scales, such small changes may accumulate, resulting in appreciable changes of mRNA abundance, whereas protein abundance would remain roughly constant. This model is a generalization of the concept of compensatory mutations that explains the rapid divergence of some *cis*-regulatory nucleotide sequences despite the maintenance of stable transcript levels [39], or the conserved expression of assembled protein complexes despite variable expression patterns of their individual components [40].

The presence of several interacting levels of protein-abundance regulation also would explain another two of our observations: a wide variance of the number of mature proteins per transcript, and a correspondingly low correlation between protein and transcript abundance within an organism (interestingly, the latter correlation is quite similar between our *C. elegans* data and data published in yeast [41] [$R_S = 0.57$]). Our data, in principle, provides an opportunity to study transcript features that would directly influence the ratio of proteins per transcript (and thereby potentially uncover novel mechanisms of translational regulation). However, when checking the influence of transcript length, GC content, or UTR length, we failed to detect correlations with protein/transcript ratios (unpublished data). We did observe a weak, but significant, positive correlation of our protein/transcript ratios and experimental protein half-life measurements of orthologous proteins in yeast [42] (unpublished data), suggesting that protein stability is indeed one of the factors determining the steady-state protein/transcript ratio.

We note that the most abundant proteins (often found in central pathways like energy metabolism or protein synthesis) also tend to be the ones that show the best abundance correlation between species. This may simply be the case because of a greater relative measurement accuracy for abundant proteins. However, highly expressed genes are also more likely to be housekeeping genes [43], and may thus be more likely to be under the same evolutionary pressures in different organisms. Strong and constant stabilizing selection is also consistent with our observation that amino acid sequences of more highly expressed proteins evolve more slowly (Figure 5C), mirroring the analogous observation for mRNA expression data [44].

When we stratify proteins by functional categories, we find that those involved in translation and in core metabolism are those with the most highly correlated abundances across species. These functional groups are also those where the coexpression between pairs of transcripts is most highly conserved across species [45]. Furthermore, the same categories also tend to show the best correlation *within* each organism, with respect to rank-correlation between transcripts and proteins (Table S2). We also find that the correlation between transcript and protein levels is particularly poor for genes that are presumably heavily regulated (the categories "signal transduction" or "transcriptional regulation"), arguing for abundant posttranscriptional regulation in these functional classes.

Proteins differ not only in their mean abundance, but also in the variance of this abundance among individuals ("noise") [46]. Interestingly, whereas yeast proteins involved in translation also show low levels of noise [47], other groups of proteins found here to be conserved in their abundance between species (e.g., protein metabolism) are characterized by high protein expression noise [47]. Thus, it appears that abundance fluctuations on short time scales (within populations) are partially decoupled from fluctuations on long time scales (between species). However, as natural variation is the substrate of evolutionary change, we expect that changes in mRNA levels via compensatory mutations may occur faster in proteins that exhibit higher levels of noise; this remains to be tested in future studies.

Our comparative analysis underlines clearly the necessity and usefulness of quantitative proteome analyses, since these better reflect the abundance of the actual effectors of biological processes. Most likely, the actual conservation of protein levels is even higher than what we report here, due to the shortcomings of a simple spectral-counting procedure. In fact, comparisons across organisms might generally provide a good test scenario to improve spectral-counting algorithms or other proteomics algorithms: the higher the abundance correlation, the more precise the measurements (due to the high number of data points, and due to the quickly changing positions of tryptic cleavages, this is difficult to "over-train" by choosing biased parameters). With respect to the transcriptomics datasets that we used, the above test argued for a better quality of the Affymetrix data, as compared to SAGE, because the latter were seen to correlate less well across organisms. This is intriguing, and it may point to additional biases in the SAGE procedure (for example, due to the added molecular biology steps of cleavage and ligation) [48].

For those instances where orthologs were *not* found to be of similar abundance, one can speculate that this difference reflects differing roles (or even molecular functions) of the orthologs. Thus, these proteins are of particular interest when studying the evolutionary differences between species. Alternatively, differences in technical aspects for particular proteins might occur, such as shifted or absent trypsin cleavage sites or differences in protein solubility. Interestingly, we did not lose the observed interspecies correlation even for quite low-abundance proteins such as those involved in signal transduction (our measurements have a dynamic range of more than three orders of magnitude). This means that low-abundance measurements are still quantitative, at least to some degree.

In our analysis of gene families with lineage-specific duplications, we found that duplicated proteins generally have lower abundance than their nonduplicated counterparts, whereas the summed abundances per gene family remained roughly constant. This finding might be most parsimoniously explained by a prevalence of subfunctionalization among duplicated genes, although it is also consistent with other scenarios (e.g., complementarity of tissue expression domains, functional fine-tuning, or subfunctionalization followed by neofunctionalization [49]). Of course, protein abundances alone cannot directly inform us about any changes in the functions of duplicated genes. However, our finding does suggest that cases where an increased demand for protein product would provide the sole driving force behind gene copy retention are probably rare.

With our dataset, we established an inventory of where and how proteins of interest can be specifically accessed using MS. It enables the generation of a proteotypic peptide library (i.e., peptides in a protein sequence that are most likely to be consistently and confidently observed by current MS-based proteomics methods). This library in turn can be used for targeted analyses and comparative studies of expressed proteins [10,50–52] by spiking the samples to be analyzed with chemically synthesized proteotypic peptides, or by selected reaction monitoring (SRM) MS. Our *C. elegans* proteome dataset will be made publicly available within WormBase and will thus be useful for the entire *C. elegans* research community. In general, proteomics data like ours is closer to the biologically active players than transcriptomics data. It should therefore be increasingly used to investigate biological phenomena and mechanisms underlying disease pathogenesis such as neuronal degeneration and cancer development, and for the identification of conserved therapeutic target proteins.

Materials and Methods

C. elegans. *C. elegans* wild-type strain N2 (Bristol) was grown on 9-cm nematode growth medium (NGM) agar plates seeded with a lawn of the *E. coli* strain OP50 or in 100 ml of liquid cultures in S-basal buffer in beveled flasks. Worms were harvested from plates or liquid culture, and separated from the bacteria by washing with water or sucrose flotation. For the collection of embryos, the worms were synchronized, and eggs were removed from agar plates or obtained from the hermaphrodites by bleaching. Worm and egg pellets were homogenized with glass beads (diameter of 212–300 μm; Sigma-Aldrich) in the ratio of 1:1:2 (worms:beads:buffer) in a cell disrupter (FastPrep FP120, Thermo Savant; Qbiogene) at 4 °C three times for 45 s at level 6. The buffer used was 50 mM Tris/HCl (pH 8.3), 5 mM EDTA, 8 M urea. After glass bead treatment, 0.125% SDS was added, and the homogenate was incubated for 1 h at room temperature (RT) to solubilize proteins. For other experiments, the worms or eggs were homogenized with glass beads in 50 mM Tris/HCl (pH 8.3), 5 mM EDTA, then 0.75% or 1% Rapigest (Waters) was added, the homogenate was heated at 95 °C for 5 min, and incubated at RT for 30–60 min with gentle agitation. Cell debris was removed by centrifugation, and the protein concentration was determined using the Bradford reagent (Sigma-Aldrich).

Tandem mass spectrometry. The peptides were subjected to reversed-phase capillary chromatography using a 75-μm × 8-cm self-packed C18 column (Magic C18; Michrom) at a flow rate of 250 nl/min. Peptides were eluted with a gradient between solvent A (5% ACN, 0.2% formic acid) and solvent B (80% ACN, 0.2% formic acid). The gradient was from 5% up to 45% solvent B within 69 min. The peptides were identified by CID (collision induced dissociation) on a Thermo-Finnigan ion trap mass spectrometer "LTQ". Six dependent scans followed each survey scan. Raw data were converted into mzXML files and searched against a *C. elegans* database derived from the Wormpep database (http://www.wormbase.org, release WS140) using the Sequest program [53]. The search parameters used were two missed cleavage sites, two tryptic termini, a mass tolerance of 3 Da for the parent ion and 0.95 Da for the fragment ion, optional oxidized methionine, and depending on the experiment, modified cysteine. Peptide assignments were statistically validated at peptide level using PeptideProphet [54], and peptides with a probability score of 0.9 or higher and the proteins they belong to were selected. For the qualitative analysis of the proteome (Figure 2), peptides matching to more than one protein (such as duplicated tubulins or histones), or matching to several splice variants of a protein, were counted only once (for the first entry of the search results). For the quantitative analysis, however, such peptides were assigned fractionally (see below). From a total of 18 different experiments (Table S3), we identified 10,977 proteins from 10,631 gene loci (Table S1). The comparative analysis of the different protein parameters was also based on WS140. For technical reasons, all the information for the other functional analyses was extracted from release WS160 using WormMart (http://www.wormbase.org/biomart/martview). The FDR for single hits was estimated first based on an experiment in which isoelectric focusing of peptides was performed on an immobilized pH gradient strip (pH range 3–5.6), followed by subsequent analysis of computationally predicted pIs for each peptide identification, and second by a new model based on a decoy search strategy (L. Reiter, M. Claassen, S. P. Schrimpf, J. M. Buhmann, M. O. Hengartner, et al., unpublished data). To evaluate potential bacterial contamination in our dataset, one experiment was searched against a combined *C. elegans* (WormBase WS140) and *E. coli* (SPproteomes at the European Bioinformatics Institute [EBI], release 2005-03-19, 4,338 entries) database using the same search parameters as for the searches against the *C. elegans* database.

Bias analysis of protein parameters. After redundancy analysis, 22,269 distinct proteins (including splice variants, WormBase WS140) and 10,977 proteins in our dataset were compared for the bias analysis with respect to different protein parameters. Tools from the ExPASy Web site (http://www.expasy.ch) were used to calculate the pIs of proteins (protein parameter tool "protparams") and their hydrophobicity (gravy computation "grand average hydrophobicity"). The statistical analysis shown in Figure S8 was carried out as described before [10]; the *p*-values for all parameter analyses were 1E–10 or better.

Transmembrane domains and GO slim terms. The number and orientation of transmembrane domains of the proteins in WormBase (WS160) and in our dataset were predicted using Phobius [11]. Only gene loci—not splice variants—were processed. Whenever transmembrane predictions differed for splice variants, the predictions for the longest splice variants were used. For the GO slim analysis, the GO terms listed in WormBase (WS160) were mapped onto higher-level terms using the GO slim guide (http://www.geneontology.org), with two exceptions: the terms "membrane" and "integral to membrane" were not mapped to the higher category term "cell," but instead were retained. In Figures 2E and S9, we assigned the GO slim terms of the category "molecular function" to the predicted transmembrane proteins. For 412 proteins, there was more than one entry for molecular function. For the statistical analysis of the GO slim categories in Figure 2, we applied the Fisher exact test and included the Bonferroni correction for multiple testing. We plotted the log ratio of observed versus expected, using the proportions in WormBase as the expectation. The GO slim categories with a *p*-value better than 0.05 are shown (Figure 2E and 2F).

Genome annotation. We mined our dataset for nonannotated translated regions by preparing a whole-genome open reading frame database that was searched using the Sequest algorithm [53]. To do this, WormBase release WS160 was used to translate each chromosome into all six reading frames. Open reading frames longer than 20 amino acids were assembled into a database with headers containing the coordinates of the sequences on the genome. The resulting database contains 3,136,258 open reading frames and 132,018,220 amino acids. A subset of our data (experiment 15) obtained by isoelectric focusing, comprising approximately 304,000 MS/MS spectra, was searched at the Functional Genomics Center Zurich. We allowed fully tryptic peptides with up to two missed cleavages, and specified oxidized methionine as variable modification and carbamylated cysteine as static modification. The results were further analyzed with PeptideProphet [54], and 27,940 search hits with a PeptideProphet score greater than or equal to 0.95 were selected. From these, we removed 26,952 scans that also generated a hit against the normal Wormpep140 protein database with a score greater than 0.8. Of the remaining 988 spectra, 789 were further observed to exist in Wormpep178 or an *E. coli* database and were therefore omitted, resulting in a final set of 199 spectra belonging to 173 different peptides. For the resulting peptides, a theoretical pI value was calculated and compared to the mean pI of all peptides in the

corresponding fraction. Only peptides with a delta pI value smaller than or equal to 0.5 were selected. This resulted in 78 distinct peptides.

Operons. WormMart (http://www.wormbase.org/biomart/martview) was used to extract operon architectures from WormBase release WS160. To test whether the coregulation of genes in operons would be detectable also at the level of translated proteins, operons were first divided into length classes (here, length is defined as the number of cotranscribed genes in each operon). For each length class, the fraction of operon genes was then determined for which at least one peptide was detected in at least one proteomics experiment. This fraction determines how many proteins should, on average, be detectable from a single operon if expression of the operon genes were truly independent (when assuming independence, the number of detections per operon should follow a binomial distribution, shown as grey lines in Figure S2). Applying the two-sided Kolmogorov-Smirnov test yielded *p*-values better than 1E–10. For the study of operons in specific stages (Figure S3), only the proteome data was analyzed, limited to the experiments done in these stages (with concomitantly reduced spectral counts).

Semiquantitative interspecies proteomics comparison. For the semiquantitative comparison between *C. elegans* and *D. melanogaster* proteomes, we used the STRING database and the Smith-Waterman similarity relations stored therein to compute orthologous groups [55]. This analysis retrieved 4,184 loci in *C. elegans*, and 4,302 in *D. melanogaster*. When working with orthology sets, each pair of orthologs was aligned with "muscle" [56], available from http://www.drive5.com/muscle/. The protein sequences used were extracted from WormBase WS160 and from FlyBase release 5.1 (http://flybase.org). Due to lineage-specific gene duplications, some proteins had several orthologs. For the interspecies abundance correlation comparison, we summed up the abundances in these cases.

We independently tested another source of orthology information, InParanoid [57], which resulted in slightly more orthologs but also in a somewhat lower interspecies abundance correlation ($R_S = 0.76$ versus 0.79). Conversely, we also tested a stricter set of orthologs, to test for and exclude artifacts caused by potentially undetected paralogy. To conclusively separate orthologs from paralogs can be difficult, and this is the subject of intense study [58–60]. Therefore, we constructed a very strict set of orthologs by searching for reciprocal best matches between worm and fly, with the additional constraint that any extra homologs within these genomes had to exhibit no more than half the alignment score than the score *between* these organism (plus, the score between the organisms had to be 60 bits or higher). This strict set contained only 2,001 pairs of orthologs, and resulted in an interspecies abundance correlation of 0.80. This shows that our high correlation is not caused, or affected, by the presence of paralogs in the comparison.

We calculated the relative abundance of a protein by counting how often any of its amino acids had been identified in any peptide, divided by the total number of amino acids of the protein sequence. A length restriction to peptides with ≥ 7 and ≤ 40 amino acids (modified from [22]) was applied.

$$a = \frac{\sum_i number(p_i) \cdot length(p_i)}{\sum_j length(q_j) \cdot f(q_j)}$$

where a = protein abundance, p = identified peptides, q = tryptic peptides (in silico digest), and $f(q)$ = peptide length correction factor.

The peptide-length correction factor takes into account the technical bias of the MS instrument, which resulted in certain peptide lengths being observed more often than others. This was learned from the data by comparing the observed peptide-length spectrum with the expected, and was corrected accordingly (similar to [22]). In our hands, peptide length proved to be the most important determinant of peptide observability, since using the original APEX implementation ("absolution protein expression profiling") [22] or a retrained version of the same classifier, did not further improve the observed cross-organism abundance correlation between *C. elegans* and *D. melanogaster* ($R_S = 0.78$).

A relative protein abundance of 1 means that the total number of amino acids in the identified peptides equals the number of amino acids in the protein. Whenever a peptide could be assigned to several proteins (because of identical predicted tryptic peptides), the amino acids were assigned fractionally. Peptides specific for any of the splice isoforms originating from a given locus were pooled. This approach means that the unit of interest in our comparisons is the gene locus—not individual splice isoforms—consistent with the observed lack of conservation of alternative splicing at very large evolutionary distances [61]. Finally, protein abundances were normalized to total amount of protein detected. To plot the data, orthologs were binned into eight groups of equal size (sorting for binning was $x + y$), and the means, as well as first and third quartiles, for each group were calculated. For the comparison of gene and protein expression, SAGE data for *C. elegans* [27] were downloaded from http://tock.bcgsc.bc.ca/cgi-bin/sage160. In order to best reflect the developmental stages analyzed in our proteome data, we chose the stages SWN21, SWL12, SWL21, SWL32, SWL41, SWYA1, MIXED, SW022, and DAUER. Only entries with "source = coding_RNA" were considered, and the average of the nine columns was calculated. SAGE data for *D. melanogaster* [29] were obtained from Professor San Ming Wang (Northwestern University, Evanston, Illinois). *D. melanogaster* SAGE tags were mapped to all transcripts from FlyBase release 5.3. The *C. elegans* Affymetrix GeneChip data were obtained from the Genome British Columbia *C. elegans* Gene Expression Consortium at http://elegans.bcgsc.bc.ca. The *D. melanogaster* Affymetrix GeneChip data [28] were obtained from http://www.flyatlas.org. For 2,695 pairs of orthologs protein abundance, SAGE and Affymetrix data were compared (in case of several paralogs, only one of them had to have data from all three measurements). For the comparisons of different abundances, Spearman rank correlation coefficients were computed to avoid assumptions about the underlying distributions. Probabilities for the correlation coefficients were calculated as implemented in R; all corresponding *p*-values were better than 2.2E–16. Further supporting the validity of spectral counting as a semiquantitative measure is a comparison of *C. elegans* protein abundance data against protein abundance data in yeast [41]. Importantly, the latter is *not* based on MS, but on immunodetection of tagged open reading frames. Orthologs correlate linearly in their abundance over two orders of magnitude ($R_S = 0.54$; Figure S10). The correlation for sequence conservation (aligned to *D. melanogaster*) and protein abundance was calculated for 4,013 *C. elegans* proteins. Orthologs were binned into eight groups of equal size (Figure 5C).

Supporting Information

Figure S1. Tandem Mass Spectra of Novel Peptides

The annotated MS/MS spectra of peptides from (A) T01B11.3 (SYN-4) and (B) F47B7.7.

Found at doi:10.1371/journal.pbio.1000048.sg001 (289 KB PDF).

Figure S2. Coordinated Expression of Operon Genes

The number of detected loci per operon deviates from what would be expected under simple independence, as shown exemplary for operons of lengths 4–6 (A–C). A higher fraction of operons than expected is either fully expressed (all proteins detected) or hardly expressed at all (none or only few proteins detected).

Found at doi:10.1371/journal.pbio.1000048.sg002 (23 KB PDF).

Figure S3. Proteins Encoded by Operon Genes Are More Abundant Than Those of Singleton Genes, Even When Focusing Exclusively on Embryos, L1, and Dauer Larvae

Although clearly significant, the effect size is lower than for the whole, presumably due to undersampling (each plot represents less than 12% of the total data). Medians are indicated as black dots, and whiskers encompass the range from 25% to 75% of values.

Found at doi:10.1371/journal.pbio.1000048.sg003 (775 KB PDF).

Figure S4. Comparing the Abundances of Proteins and Transcripts, Specifically in Embryos Only (Worm and Fly)

(A) Protein abundances of 1,195 conserved pairs of orthologs, which were detected in embryos of both *D. melanogaster* and *C. elegans*, and for which transcript data were available (see below). Protein abundances were estimated by spectral counting (limited to data from experiments using embryos, reducing the data to about one tenth of the total).
(B) Spearman rank correlation coefficients. Protein abundances correlate better across organisms than transcript abundances, and better than protein versus transcript within organisms.
(C) Transcript abundances of the same 1,195 conserved pairs of orthologs as in (A), from published measurements using Affymetrix arrays. Raw CEL files were reanalyzed using the MBEI algorithm as implemented in the cCHIP package. *C. elegans* embryo data were from the Genome British Columbia *C. elegans* Gene Expression Consortium, and the *D. melanogaster* data was from the ArrayExpress database,

using wild-type controls from the experiments E-GOED-2780, E-MEXP-879, and E-MEXP-623, which cover embryonic development at a number of time points ranging from 2.5 h to 19 h after egg-laying. Medians of equal-sized bins are indicated as crosses; whiskers encompass the range from 25% to 75% of values.

Found at doi:10.1371/journal.pbio.1000048.sg004 (430 KB PDF).

Figure S5. Down-Sampling of Proteomics Data to Ensure Independence of Peptide Counts

Individually aligned pairs of orthologs were scanned for residues R and K, in order to identify aligned tryptic cleavage sites (red vertical lines). Peptide identifications were then down-sampled in alternating stretches of the alignment, to make sure that orthologous peptides are counted for one of the two organisms only. The Spearman rank correlation dropped to 0.68. Intriguingly, this result is almost identical to what is expected simply due to the reduction of the data by half ($R_S = 0.67$ when randomly discarding 50% of the peptides); this shows that the strong correlation between *C. elegans* and *D. melanogaster* is not simply due to a tendency of orthologous peptides to be detected equally well. To also exclude local effects (i.e., dependencies between neighboring peptides), an independent test was performed for which proteins were cut in half, and N-terminal and C-terminal fragments were counted separately. In this test, when comparing orthologous proteins only via nonoverlapping halves (N-terminus versus C-terminus), the cross-organism correlation dropped to 0.66. In contrast, when comparing N-termini with N-termini (or C-termini with C-termini), the correlation was higher (0.71). This indicates that there are indeed some local dependencies between peptide counts, but not enough to explain the high interorganism correlation we observe when using the full data.

Found at doi:10.1371/journal.pbio.1000048.sg005 (104 KB PDF).

Figure S6. Expression Levels of Duplicated Genes

Genes were classified as duplicated when an orthologous group contained more than one gene in one lineage, but only a single gene in the other lineage. Abundances of duplicated genes were either plotted separately (A), or pooled for each group (B). Columns marked with asterisks (***), are significantly different (*p*-value better than 1E–15). Medians are indicated as black dots, and whiskers encompass the range from 25% to 75% of values.

Found at doi:10.1371/journal.pbio.1000048.sg006 (227 KB PDF).

Figure S7. Fly Orthologs of Worm Operon Genes

D. melanogaster genes were classified according to whether their orthologs in *C. elegans* are part of operons. Note that these genes are not organized in operons in the fly, nor are they even neighbors on the chromosome. Still, fly proteins are more abundant when their worm orthologs are arranged in operons. *p*-values: a single asterisk (*) indicates better than 1E–5; double asterisks (**) indicate better than 1E–10; and triple asterisks (***) indicate better than 1E–15. Medians are indicated as black dots, and whiskers encompass the range from 25% to 75% of values.

Found at doi:10.1371/journal.pbio.1000048.sg007 (174 KB PDF).

Figure S8. Statistical Bias Analysis of the Protein Parameters Length, pI, and Hydrophobicity

Distributions of the parameters of the identified proteins versus all proteins in WormBase (WS140). Overrepresented areas are shown in green, underrepresented areas in yellow (*p*-values were better than 1E–10; for details about the applied statistics, see [10]).

Found at doi:10.1371/journal.pbio.1000048.sg008 (19 KB PDF).

Figure S9. The Predicted *C. elegans* Transmembrane Proteome and Its Molecular Function

We predicted the transmembrane topology of the entire *C. elegans* proteome and included the molecular function of the proteins with transmembrane helices. The percentages are referring to the entire dataset. Proteins with a cytoplasmic C-terminus were plotted upwards; proteins with an extracytoplasmic C-terminus were plotted downwards. The color code for the molecular function is indicated.

Found at doi:10.1371/journal.pbio.1000048.sg009 (1.39 MB PDF).

Figure S10. Further Support for the Validity of Protein Quantification in *C. elegans*, from Comparison against Published *S. cerevisiae* Data

Protein abundances deduced from spectral counting (*C. elegans*) and from protein tagging and immunodetection (yeast [41]) of 1,092 pairs of orthologs from both species yielded a correlation coefficient of $R_S = 0.54$. Medians of equal-sized bins are indicated as crosses; whiskers encompass the range from 25% to 75% of values.

Found at doi:10.1371/journal.pbio.1000048.sg010 (143 KB PDF).

Table S1. Identified *C. elegans* Proteins and Peptides

In our shotgun proteomic approach, 84,962 unique peptides were identified after filtering with the PeptideProphet probability score equal to or greater than 0.9. The scan numbers, the peptides, and the coding sequence of the proteins they mapped to are listed.

Found at doi:10.1371/journal.pbio.1000048.st001 (8.55 MB ZIP).

Table S2. Intraspecies Protein versus Transcript Correlations, Broken Down into Functional Categories

Both fly and worm proteins were mapped to GO slim categories by a similar procedure. In both organisms, comparable categories show a high or low correlation. In addition, even categories of relatively low abundance (e.g., "DNA metabolism") can have a high correlation, indicating that the ranking is not simply based on measurement accuracy.

Found at doi:10.1371/journal.pbio.1000048.st002 (39 KB PDF).

Table S3. List of Experiments

The experiment ID, the developmental stages of the worm, the sample type, and the biochemical separation methods are listed.

Found at doi:10.1371/journal.pbio.1000048.st003 (18 KB PDF).

Acknowledgments

We thank Frank Potthast and Christian Panse for the database searches; Bernd Roschitzki, Mike Scott, Bertran Gerrits, and René Brunisholz for technical support; Ralph Schlapbach for access to the Functional Genomics Center Zurich; and San Ming Wang for providing SAGE tags for *D. melanogaster*.

Author contributions. SPS conducted the majority of the proteomics experiments. MW did part of the analyses for functional characterization of the dataset. LR administrated the proteome dataset. CHA did the bias analyses. MJ, JM, and PEH helped with selected proteomics experiments. EB and SM generated the *D. melanogaster* proteome dataset. MJL helped with the functional analyses. RA coinitiated the project. CvM performed the operon analysis and supported MW with the functional analyses. SPS, CvM, MJL, and MOH wrote the manuscript; and MOH initiated and supervised the whole project.

Funding. This work was funded by the University of Zurich Research Priority Program in Systems Biology/Functional Genomics, the Swiss National Science Foundation, the GEBERT RÜF Foundation, SystemsX, and the Ernst Hadorn Foundation. MJ and LR were supported by a grant from the Research Foundation of the University of Zurich. MJ was also supported by a fellowship from the Roche Research Foundation. JM was supported by a fellowship from the Swedish society for medical research (SSMF). The *C. elegans* SAGE data were produced at the Michael Smith Genome Sciences Centre with funding from Genome Canada. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript

Competing interests. The authors have declared that no competing interests exist.

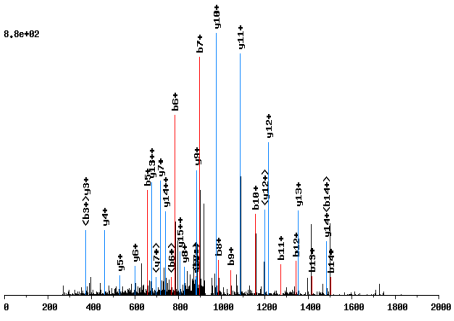
References

1. O'Brien KP, Westerlund I, Sonnhammer EL (2004) OrthoDisease: a database of human disease orthologs. Hum Mutat 24: 112–119.
2. C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282: 2012–2018.
3. Anderson L, Seilhamer J (1997) A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18: 533–537.
4. Greenbaum D, Colangelo C, Williams K, Gerstein M (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol 4: 117.
5. Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19: 1720–1730.
6. Schrimpf SP, Langen H, Gomes AV, Wahlestedt C (2001) A two-dimensional protein map of Caenorhabditis elegans. Electrophoresis 22: 1224–1232.
7. Mawuenyega KG, Kaji H, Yamuchi Y, Shinkawa T, Saito H, et al. (2003) Large-scale identification of Caenorhabditis elegans proteins by multi-

dimensional liquid chromatography-tandem mass spectrometry. J Proteome Res 2: 23–35.
8. Merrihew GE, Davis C, Ewing B, Williams G, Kall L, et al. (2008) Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. Genome Res 18: 1660–1669..
9. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, et al. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17: 994–999.
10. Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, et al. (2007) A high-quality catalog of the Drosophila melanogaster proteome. Nat Biotechnol 25: 576–583.
11. Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338: 1027–1036.
12. Kim H, Melen K, Osterberg M, von Heijne G (2006) A global topology map of the Saccharomyces cerevisiae membrane proteome. Proc Natl Acad Sci U S A 103: 11142–11147.
13. Daley DO, Rapp M, Granseth E, Melen K, Drew D, et al. (2005) Global topology analysis of the Escherichia coli inner membrane proteome. Science 308: 1321–1323.
14. Troemel ER, Kimmel BE, Bargmann CI (1997) Reprogramming chemotaxis responses: sensory neurons define olfactory preferences in C. elegans. Cell 91: 161–169.
15. Colosimo ME, Tran S, Sengupta P (2003) The divergent orphan nuclear receptor ODR-7 regulates olfactory neuron gene expression via multiple mechanisms in Caenorhabditis elegans. Genetics 165: 1779–1791.
16. Lans H, Jansen G (2006) Noncell- and cell-autonomous G-protein-signaling converges with Ca2+/mitogen-activated protein kinase signaling to regulate str-2 receptor gene expression in Caenorhabditis elegans. Genetics 173: 1287–1299.
17. Blumenthal T, Gleason KS (2003) Caenorhabditis elegans operons: form and function. Nat Rev Genet 4: 112–120.
18. Lercher MJ, Blumenthal T, Hurst LD (2003) Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res 13: 238–243.
19. Liu H, Sadygov RG, Yates JR 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76: 4193–4201.
20. Zybailov B, Coleman MK, Florens L, Washburn MP (2005) Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. Anal Chem 77: 6218–6224.
21. Kislinger T, Cox B, Kannan A, Chung C, Hu P, et al. (2006) Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Cell 125: 173–186.
22. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol 25: 117–124.
23. Vogel C, Marcotte EM (2008) Calculating absolute and relative protein abundance from mass spectrometry-based protein expression data. Nat Protoc 3: 1444–1451.
24. Blumenthal T (2004) Operons in eukaryotes. Brief Funct Genomic Proteomic 3: 199–211.
25. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. (2003) A protein interaction map of Drosophila melanogaster. Science 302: 1727–1736.
26. Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, et al. (2006) Genomic and functional evolution of the Drosophila melanogaster sperm proteome. Nat Genet 38: 1440–1445.
27. McKay SJ, Johnsen R, Khattra J, Asano J, Baillie DL, et al. (2003) Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans. Cold Spring Harb Symp Quant Biol 68: 159–169.
28. Chintapalli VR, Wang J, Dow JA (2007) Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet 39: 715–720.
29. Lee S, Bao J, Zhou G, Shapiro J, Xu J, et al. (2005) Detecting novel low-abundant transcripts in Drosophila. RNA 11: 939–946.
30. de Godoy LM, Olsen JV, de Souza GA, Li G, Mortensen P, et al. (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. Genome Biol 7: R50.
31. Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, et al. (2008) Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. Genome Res 18: 1133–1142.
32. Li WH, Yang J, Gu X (2005) Expression divergence between duplicate genes. Trends Genet 21: 602–607.
33. Scannell DR, Wolfe KH (2008) A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome Res 18: 137–147.
34. He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169: 1157–1164.
35. Robertson HM, Thomas JH (2006) The putative chemoreceptor families of C. elegans. WormBook Jan 6: 1–12.
36. Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, et al. (2004) A neutral model of transcriptome evolution. PLoS Biol 2: e132. doi:10.1371/journal.pbio.0020132
37. Yanai I, Korbel JO, Boue S, McWeeney SK, Bork P, et al. (2006) Similar gene expression profiles do not imply similar tissue functions. Trends Genet 22: 132–138.
38. Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, et al. (2005) The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat Genet 37: 544–548.
39. Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564–567.
40. Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. Nature 443: 594–597.
41. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in yeast. Nature 425: 737–741.
42. Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK (2006) Quantification of protein half-lives in the budding yeast proteome. Proc Natl Acad Sci U S A 103: 13004–13009.
43. Lercher MJ, Urrutia AO, Hurst LD (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet 31: 180–183.
44. Pal C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. Nat Rev Genet 7: 337–348.
45. Bergmann S, Ihmels J, Barkai N (2004) Similarities and differences in genome-wide expression data of six organisms. PLoS Biol 2: e9. doi:10.1371/journal.pbio.0020009
46. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise minimization in eukaryotic gene expression. PLoS Biol 2: e137. doi:10.1371/journal.pbio.0020137
47. Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, et al. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature 441: 840–846.
48. Wang SM (2007) Understanding SAGE data. Trends Genet 23: 42–50.
49. Hughes T, Liberles DA (2007) The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation. J Mol Evol 65: 574–588.
50. Ahrens CH, Brunner E, Hafen E, Aebersold R, Basler K (2007) A proteome catalog of Drosophila melanogaster. An essential resource for targeted quantitative proteomics. Fly 1: e1–e5.
51. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, et al. (2007) Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 25: 125–131.
52. Kuster B, Schirle M, Mallick P, Aebersold R (2005) Scoring proteomes with proteotypic peptide probes. Nat Rev Mol Cell Biol 6: 577–583.
53. Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5: 976–989.
54. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74: 5383–5392.
55. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, et al. (2007) STRING 7–recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35: D358–362.
56. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
57. Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res 36: D263–266.
58. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, et al. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5: R7.
59. Zhang P, Min W, Li WH (2004) Different age distribution patterns of human, nematode, and Arabidopsis duplicate genes. Gene 342: 263–268.
60. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39: 309–338.
61. Copley RR (2008) The animal in the genome: comparative genomics and evolution. Philos Trans R Soc Lond B Biol Sci 363: 1453–1461.
62. Keller A, Eng J, Zhang N, Li XJ, Aebersold R (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol 1: 2005 0017.

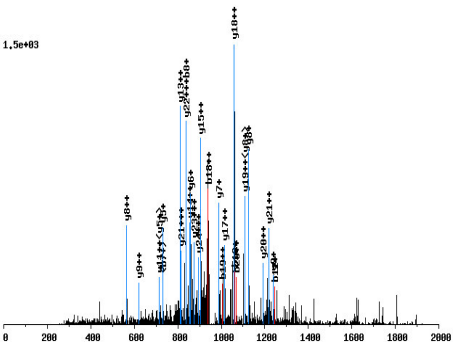
Schrimpf et al., Figure S1

A SYN-4 (T01B11.3)



b^+	#	AA	#	y^+	y^{2+}
114.1674	1	L	17		
261.3439	2	F	16	1759.9745	880.4912
390.4594	3	E	15	1612.7979	806.9029
521.6520	4	M	14	1483.6825	742.3452
658.7931	5	H	13	1352.4899	676.7489
786.9238	6	Q	12	1215.3488	608.1784
900.0832	7	I	11	1087.2181	544.1130
987.1614	8	S	10	974.0587	487.5333
1044.2133	9	G	9	886.9805	443.9942
1157.3728	10	I	8	829.9285	415.4682
1271.4766	11	N	7	716.7691	358.8885
1342.5554	12	A	6	602.6653	301.8366
1413.6342	13	A	5	531.5865	266.2972
1500.7124	14	S	4	460.5077	230.7578
1597.8291	15	P	3	373.4295	187.2187
1726.9446	16	E	2	276.3128	138.6604
	17	K	1	147.1973	74.1026

B F47B7.7



b^+	b^{2+}	#	AA	#	y^+	y^{2+}	y^{3+}
187.2211	94.1145	1	W	25			
244.2731	122.6405	2	G	24	2681.7912	1341.3996	894.6024
359.3617	180.1848	3	D	23	2624.7393	1312.8736	875.5850
430.4405	215.7242	4	A	22	2509.6507	1255.3293	837.2222
487.4924	244.2502	5	G	21	2438.5719	1219.7899	813.5292
650.6683	325.8381	6	Y	20	2381.5199	1191.2639	794.5119
749.8009	375.4044	7	V	19	2218.3440	1109.6760	740.1200
836.8791	418.9435	8	S	18	2119.2114	1060.1097	707.0758
974.0202	487.5141	9	H	17	2032.1332	1016.5706	678.0497
1061.0984	531.0532	10	S	16	1894.9921	948.0000	632.3360
1158.2151	579.6115	11	P	15	1807.9139	904.4609	603.3099
1245.2933	623.1506	12	S	14	1710.7973	855.9026	570.9377
1342.4099	671.7089	13	P	13	1623.7191	812.3635	541.9116
1443.5150	722.2615	14	T	12	1526.6024	763.8052	509.5394
1500.5669	750.7874	15	G	11	1425.4973	713.2526	475.8377
1629.6824	815.3452	16	E	10	1368.4454	684.7267	456.8204
1742.8419	871.9249	17	I	9	1239.3299	620.1689	413.7819
1879.9829	940.4954	18	H	8	1126.1705	563.5892	376.0621
2009.0984	1005.0532	19	E	7	989.0294	495.0187	330.3484
2138.2139	1069.6109	20	E	6	859.9139	430.4609	287.3099
2301.3899	1151.1989	21	Y	5	730.7984	365.9032	244.2714
2429.5206	1215.2643	22	Q	4	567.6225	284.3152	189.8794
2592.6965	1296.8522	23	Y	3	439.4917	220.2498	147.1692
2693.8016	1347.4048	24	T	2	276.3158	138.6619	92.7772
		25	R	1	175.2107	88.1093	59.0755

Figure S1
Tandem mass spectra of novel peptides. The annotated MS/MS spectra of peptides from (A) T01B11.3 (SYN-4) and (B) F47B7.7.

Schrimpf et al., Figure S2

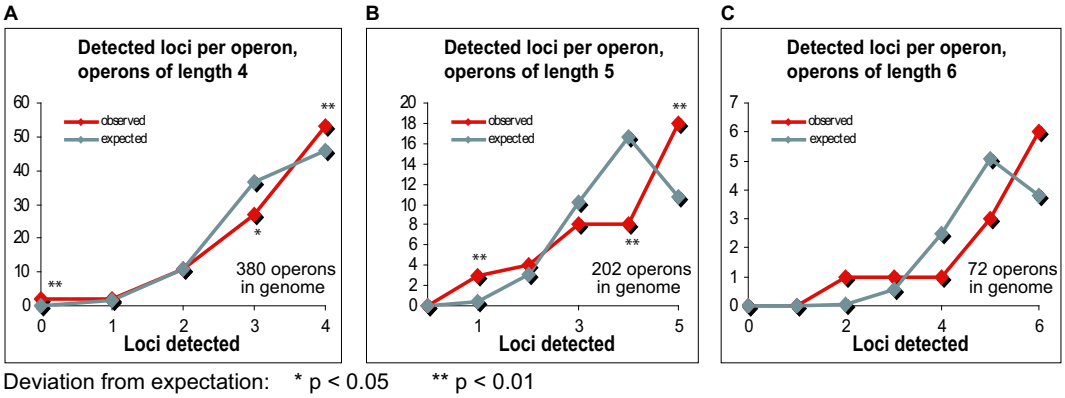


Figure S2
Coordinated expression of operon genes**.** The number of detected loci per operon deviates from what would be expected under simple independence, as shown exemplary for operons of lengths 4 - 6 (A-C). A higher fraction of operons than expected is either fully expressed (all proteins detected) or hardly expressed at all (none or only few proteins detected).

Schrimpf et al., Figure S3

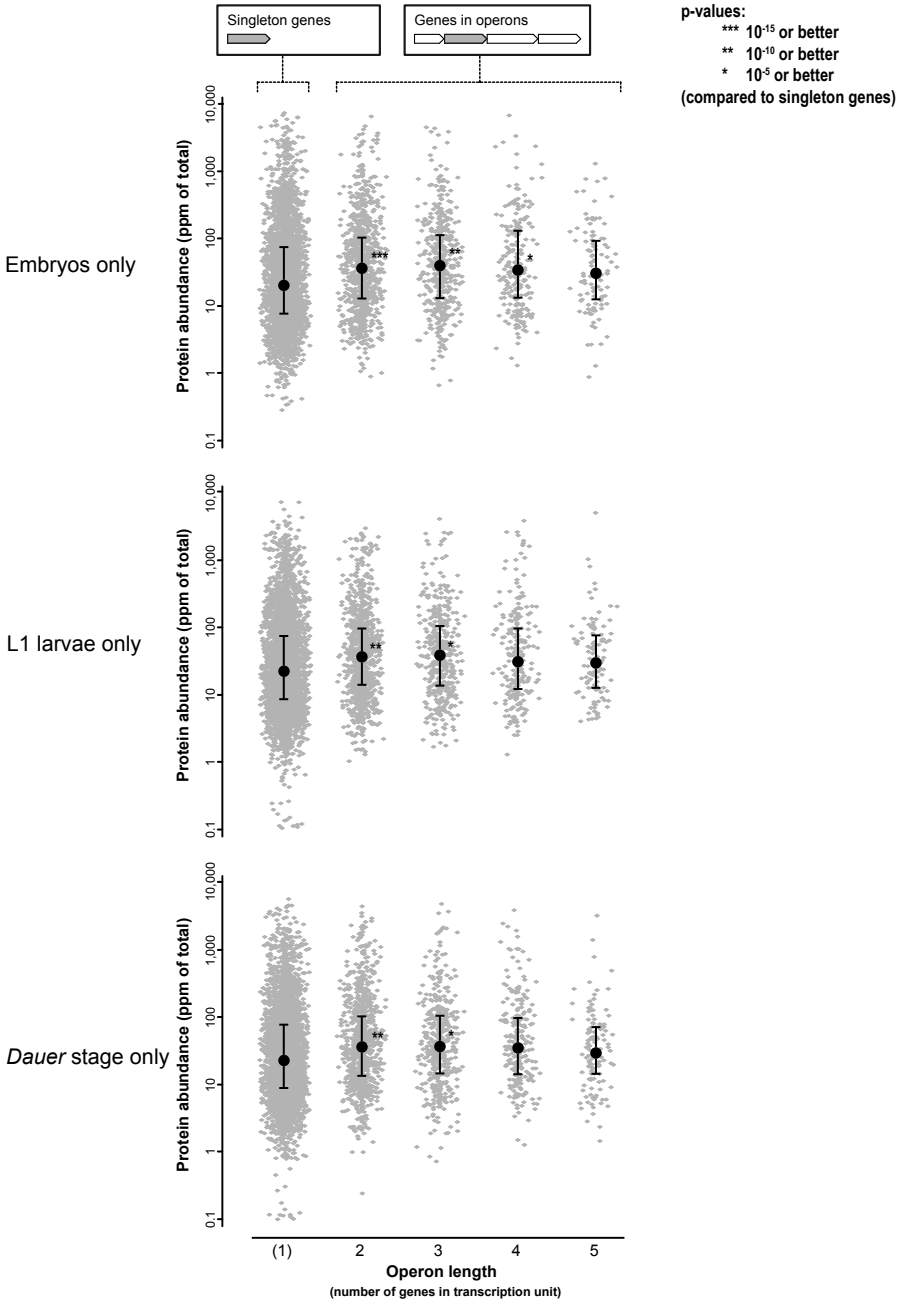


Figure S3
Proteins encoded by operon genes are more abundant than those of singleton genes, even when focusing exclusively on embryos, L1, and *Dauer* larvae. While clearly significant, the effect size is lower than for the whole, presumably due to undersampling (each plot represents less than 12% of the total data).

Schrimpf et.al., Figure S4

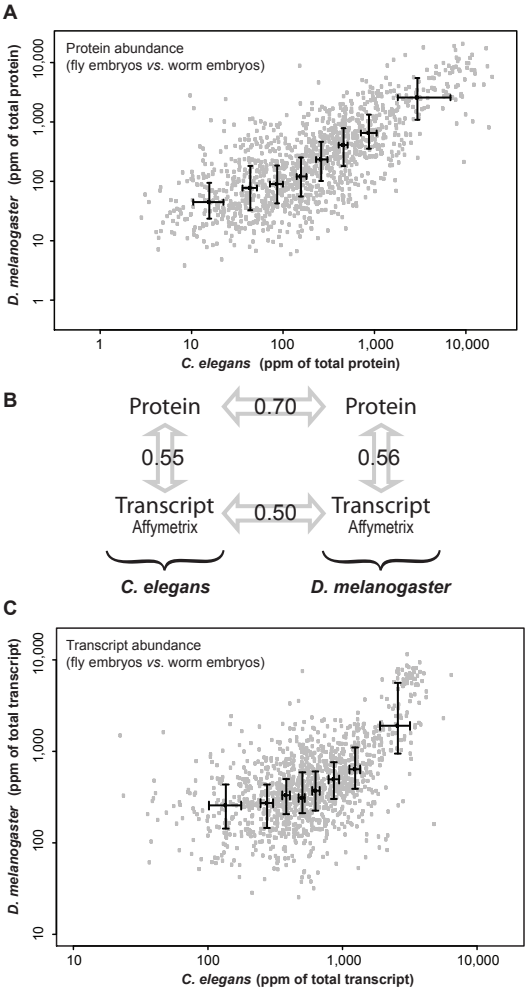


Figure S4
Comparing the abundances of proteins and transcripts, specifically in embyros only (worm and fly). (A) Protein abundances of 1,195 conserved pairs of orthologs, which were detected in embryos of both *D. melanogaster* and *C. elegans*, and for which transcript data was available (see below). Protein abundances were estimated by spectral counting (limited to data from experiments using embryos, reducing the data to about one tenth of the total). (B) Spearman's rank correlation coefficients. Protein abundances correlate better across organisms than transcript abundances, and better than protein *vs*. transcript within the organisms. (C) Transcript abundances of the same 1,195 conserved pairs of orthologs as in (A), from published measurements using Affymetrix arrays. Raw CEL-files were re-analyzed using the MBEI algorithm as implemented in the dCHIP package. *C. elegans* embryo data were from the Genome British Columbia *C. elegans* Gene Expression Consortium, and *D. melanogaster* data were from the ArrayExpress database, using wildtype controls from the experiments E-GOED-2780, E-MEXP-879, and E-MEXP-623, which cover embryonic development at a number of time points ranging from 2.5 h to 19 h after egg-laying.

Schrimpf et al., Figure S5

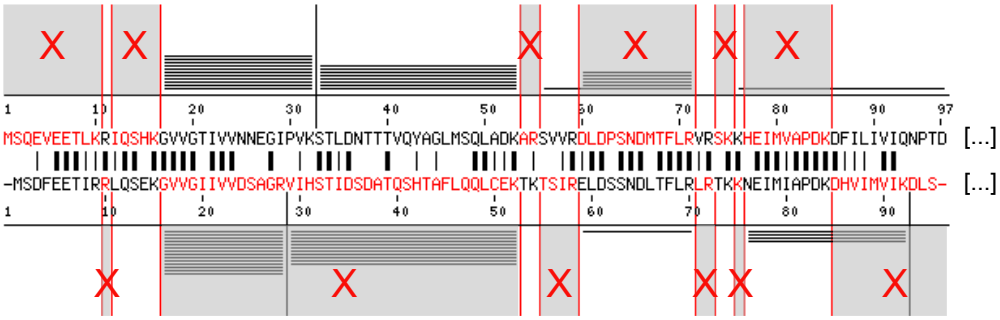


Figure S5
Down-sampling of proteomics data to ensure independence of peptide counts. Individually aligned pairs of orthologs were scanned for residues R and K, in order to identify aligned tryptic cleavage sites (red vertical lines). Peptide identifications were then down-sampled in alternating stretches of the alignment, to make sure that orthologous peptides are counted for one of the two organisms only. The Spearman's rank correlation dropped to 0.68. Intriguingly, this result is almost identical to what is expected simply due to the reduction of the data by half (R_S=0.67 when randomly discarding 50% of the peptides); this shows that the strong correlation between *C. elegans* and *D. melanogaster* is not simply due to a tendency of orthologous peptides to be detected equally well. To also exclude local effects (i.e. dependencies between neighboring peptides), an independent test was performed for which proteins were cut in half, and N-terminal and C-terminal fragments were counted separately. In this test, when comparing orthologous proteins only via non-overlapping halves (N-terminus *vs*. C-terminus), the cross-organism correlation dropped to 0.66. In contrast, when comparing N-termini with N-termini (or C-termini with C-termini), the correlation was higher (0.71). This indicates that there are indeed some local dependencies between peptide counts, but not enough to explain the high inter-organism correlation we observe when using the full data.

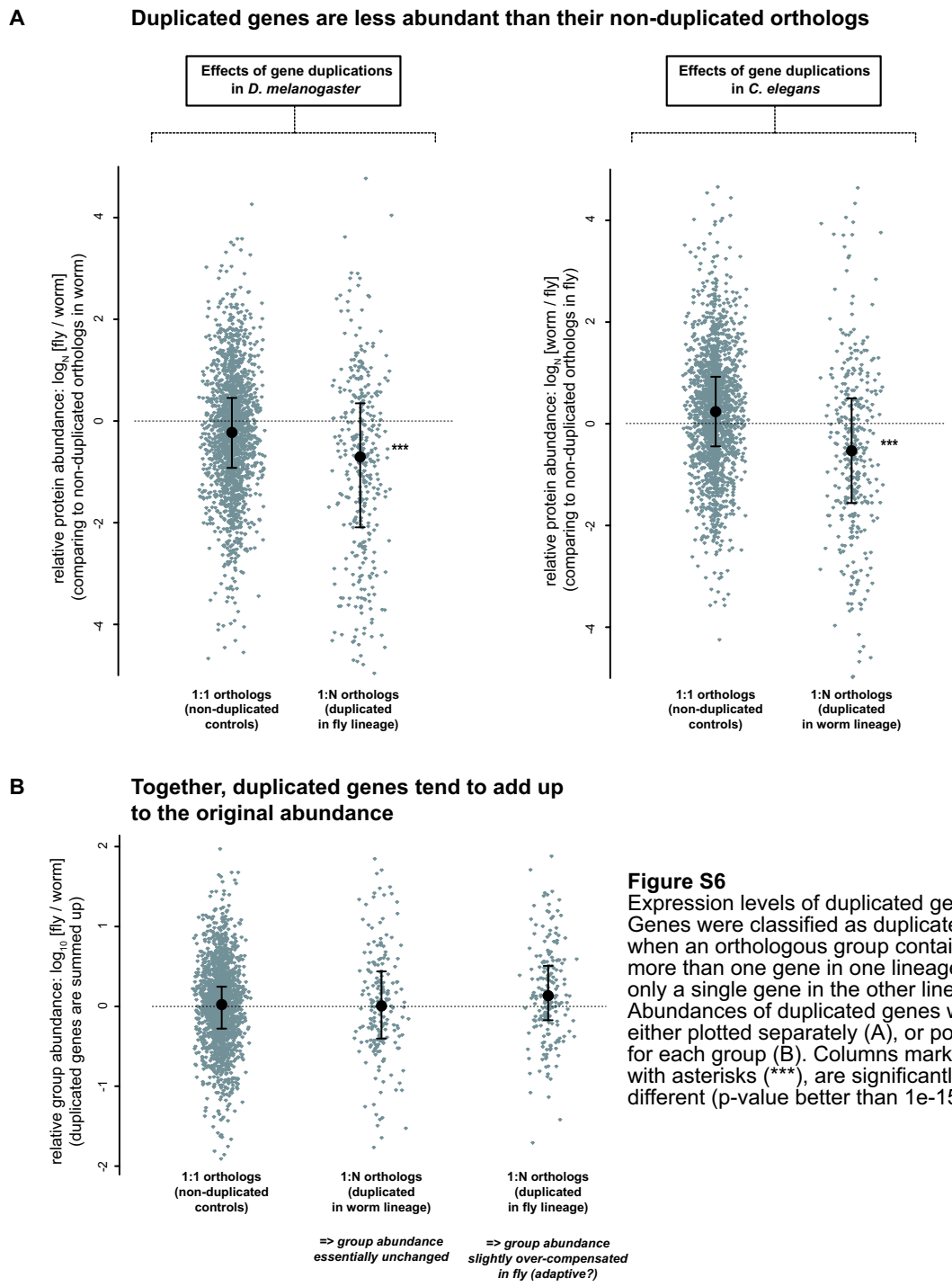


Figure S6
Expression levels of duplicated genes. Genes were classified as duplicated when an orthologous group contained more than one gene in one lineage, but only a single gene in the other lineage. Abundances of duplicated genes were either plotted separately (A), or pooled for each group (B). Columns marked with asterisks (***), are significantly different (p-value better than 1e-15).

Schrimpf et al., Figure S7

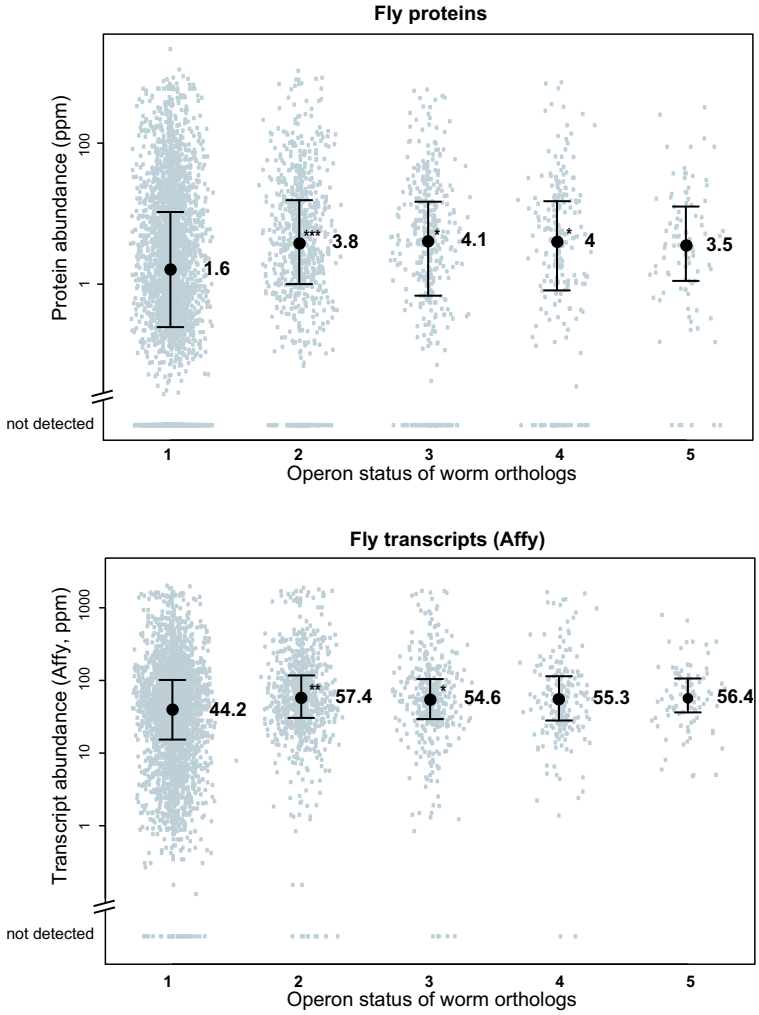


Figure S7
Fly orthologs of worm operon genes
*D. melanogaste*r genes were classified according to whether their orthologs in *C. elegans* are part of operons. Note that these genes are not organized in operons in the fly, nor are they even neighbors on the chromosome. Still, fly proteins are more abundant when their worm orthologs are arranged in operons. p-values: (**) better than 1e-10; (***) better than 1e-15.

Schrimpf et al., Figure S8

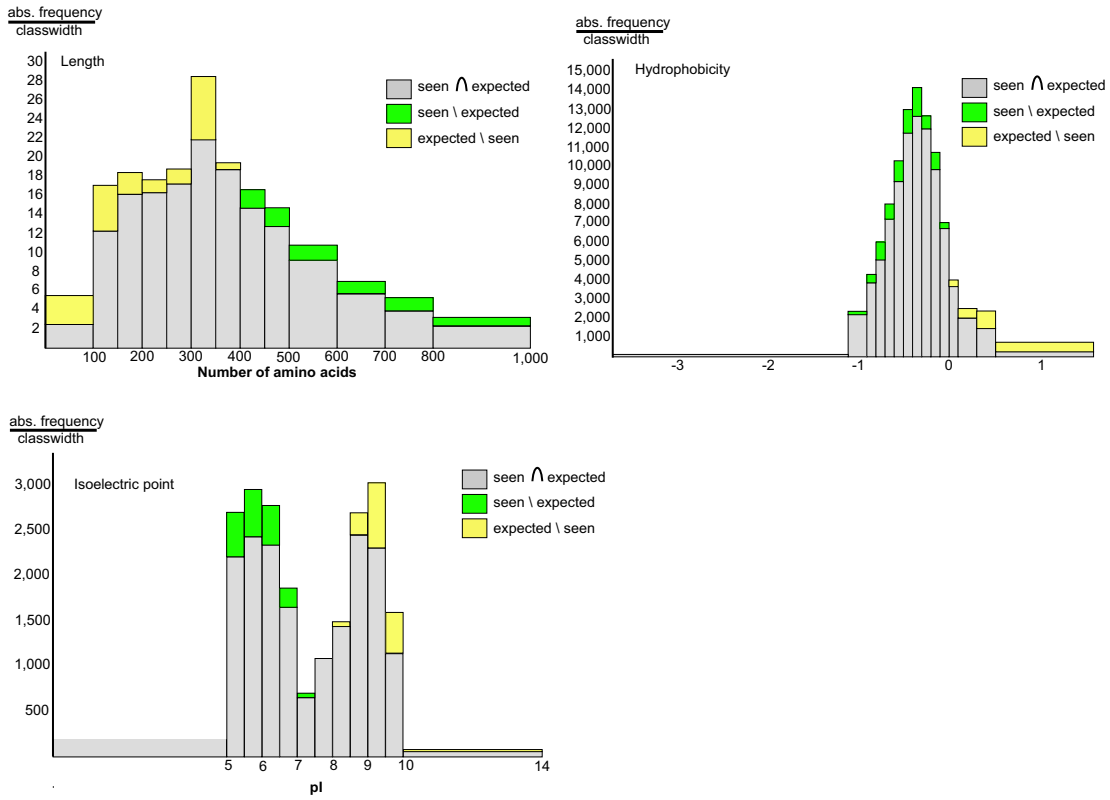


Figure S8
Statistical bias analysis of the protein parameters length, pI, and hydrophobicity**.** Distributions of the parameters of the identified proteins *vs*. all proteins in WormBase (WS140). Overrepresented areas are shown in green, underrepresented areas in yellow (p-values were better than 1e-10, for details about the applied statistics see [10]).

Schrimpf et al., Figure S9

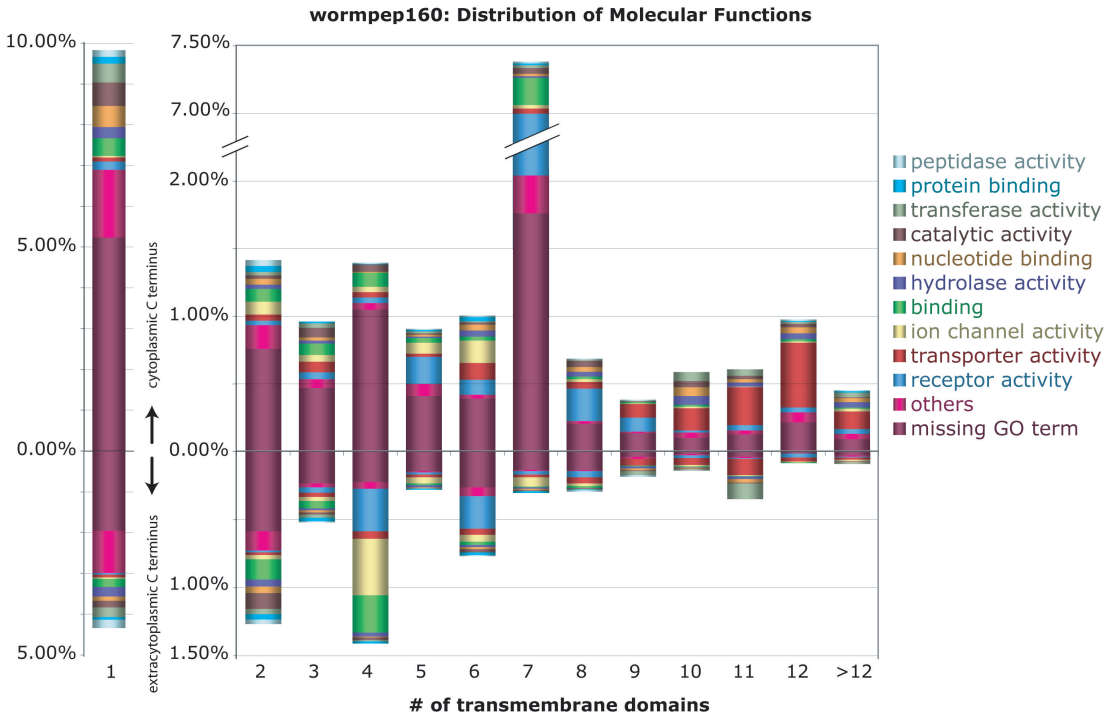


Figure S9
The predicted *C. elegans* transmembrane proteome and its molecular function. We predicted the transmembrane topology of the entire *C. elegans* proteome and included the molecular function of the proteins with transmembrane helices. The percentages are referring to the entire dataset. Proteins with a cytoplasmic C-terminus were plotted upwards, proteins with an extracytoplasmic C-terminus were plotted downwards. The colour code for the molecular function is indicated.

Schrimpf et al., Figure S10

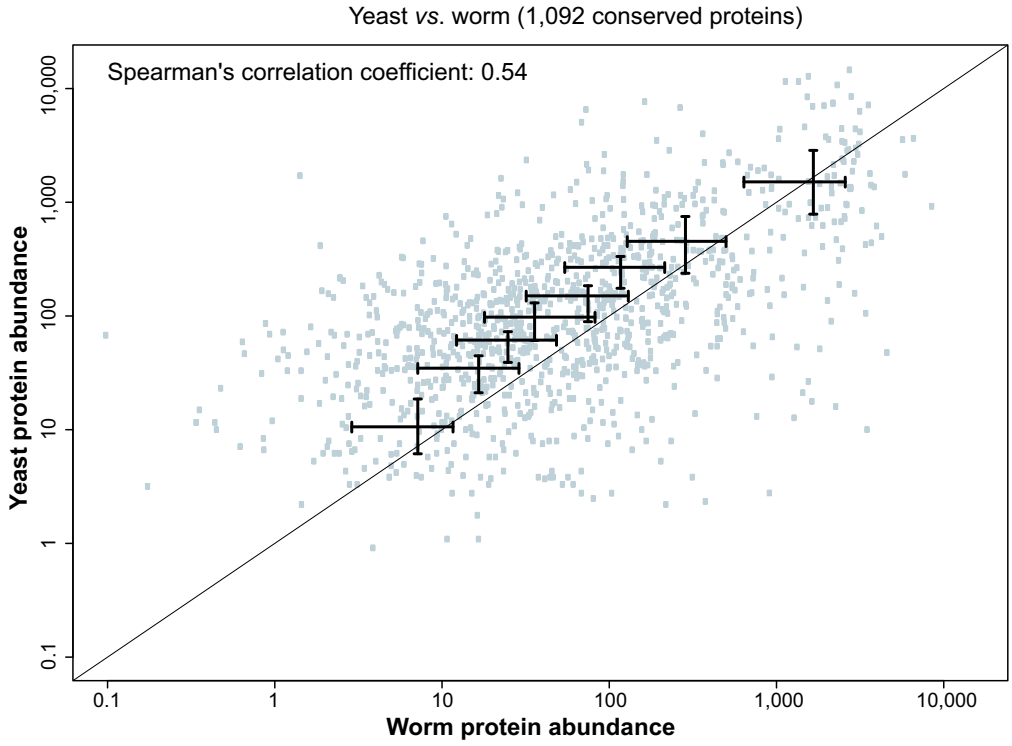


Figure S10
Further support for the validity of protein quantification in *C. elegans,* from comparison against published *Saccharomyces cerevisiae* data**.** Protein abundances deduced from spectral counting (*C. elegans*) and from protein tagging and immunodetection (yeast [41]) of 1,092 pairs of orthologs from both species yielded a correlation coefficient of R_S=0.54. Medians of equal sized bins are indicated as crosses, whiskers encompass the range from 25 to 75% of values.

5 Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome

5.1 Preface

This publication represents a continuation and extension of the work I did for the previous one (chapter 4). I did almost all the analyses, with support and many ideas from Christian von Mering, while Sabine Schrimpf and Michael Hengartner helped through discussions and feedback.

Martin Lercher contributed the generalized linear regression analysis.

Note: *I did not get the permission by Wiley-VCH to include the print edition of this publication in the electronic version of the thesis, so I included the text and figures as we submitted them.*

It was published as "Weiss et al., Proteomics 2010, 6, 1297-1306" and can be found here: http://www3.interscience.wiley.com/journal/123239141/abstract

Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome.

Manuel Weiss [1,2,3], Sabine Schrimpf [1],
Michael O. Hengartner [1], Martin J. Lercher [4] and Christian von Mering [1,2]

[1] Institute of Molecular Biology and [2] Swiss Institute of Bioinformatics, University of Zurich, Switzerland.
[3] PhD program in Molecular Life Sciences, University of Zurich and ETH Zurich, Switzerland.
[4] Dept. of Computer Science, Heinrich-Heine-University Düsseldorf, Germany.

Abstract:

Genome-wide, absolute quantification of expressed proteins is not yet within reach for most eukaryotes. However, large numbers of mass spectrometry-based protein identifications have been deposited in databases, together with information on the observation frequencies of each peptide spectrum ('spectral counts'). We have conducted a meta-analysis using several million peptide observations from five model eukaryotes, establishing a consistent, semi-quantitative analysis pipeline. By inferring and comparing protein abundances across orthologs, we observe: i) the accuracy of spectral counting predictions is increasing with sampling depth, and can rival that of direct biochemical measurements, ii) the quantitative makeup of the consistently observed core proteome in eukaryotes is remarkably stable, with abundance correlations exceeding R_S=0.7 at an evolutionary distance greater than 1000 million years, and iii) some groups of proteins are more constrained than others. We argue that our observations reveal stabilizing selection: central parts of the eukaryotic proteome appear to be expressed at well-balanced, near optimal abundance levels. This is consistent with our further observations that essential proteins show lower abundance variations than non-essential proteins, and that gene families which tend to undergo gene duplications are less well constrained than families that keep a single-copy status.

Keywords:

Protein abundance / Evolution / Shotgun Proteomics / Expression Noise / Spectral Counting / Orthologs

Introduction

Most proteins in an organism are tightly regulated in their activity – through spatial and temporal control of their production, compartmentalization in the cell, assembly into protein complexes, post-translational modifications and/or regulated degradation. Given all these levels of regulation, the molecular abundance of a protein in the cell is only one of several factors controlling its activity, and perhaps not the most important: many gene loci can tolerate copy number polymorphisms that alter their expression dosage [1, 2], there is considerable variation of protein abundances at the cell-to-cell level [3-5], and small changes in gene expression levels during evolution are often argued to be neutral, or nearly neutral [6-9].

Nevertheless, the expression levels of proteins in cells are roughly kept in line with functional requirements [10], and they can be reproducibly measured for a given set of conditions. The measured abundances can differ widely from protein to protein, typically spanning several orders of magnitude [11, 12]. The systematic biochemical measurement of absolute protein abundances in a genome-wide fashion is technically demanding, and has only been attempted in yeast [4, 12], an organism of intermediate complexity having the additional advantage of near-complete clone libraries that can provide a consistently quantifiable 'tag' at each protein. For restricted sets of proteins, or for relatively small bacterial genomes, quantitative approaches are available [13-16], for example in mass spectrometry via the addition of known amounts of mass-shifted control peptides ('spiking') or via peak-detection and -integration in ion chromatograms [17]. However, proteome-wide absolute quantification of all expressed proteins in eukaryotes remains difficult, and it has not yet been attempted for large and complex metazoans, including humans. This leaves many questions unanswered: is there a uniform optimal stoichiometry of the players in the core cellular processes across species? If so, what is this 'optimal' abundance for any given eukaryotic protein, and how stringently is it maintained by selection? Are proteins that form functional partnerships typically of the same abundance? How does intrinsic cell-to-cell variation ('noise') in protein expression translate to variability of proteome composition at evolutionary timescales?

While there are currently not enough quantitative data on proteome-wide expression in complex eukaryotes to address these questions, more *qualitative* proteomics approaches do exist, for example systematic MS-based proteome surveys of important model organisms [18-20]. These projects are typically undertaken in order to validate or correct genome annotations, to check the expression status of known or predicted genes in various tissues, and to enable the identification of non-redundant, technically suitable peptides for subsequent, more targeted routine measurements. To cover a large part of the proteome, extensive biochemical fractionation is usually necessary, making these projects large in scope and resource-intensive. Such 'shotgun proteomics' experiments generate large lists of tryptic peptide identifications, and, as a side product, information on how often the mass spectrometers have addressed any given peptide ('spectral counts'). While in most cases these data were never originally intended for use in quantification, the spectral counts do hold quantitative information, and a number of algorithms have been devised for extracting semi-quantitative abundance estimates from spectral counts [21-26].

Here, we describe an integrated analysis of published shotgun proteomics data, aiming to compare spectral counts in five eukaryotes: yeast, thale cress, human, fruit fly, and nematode (*Saccharomyces cerevisiae, Arabidopsis thaliana, Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans*). We identified gene loci that are common to all five organisms; this set of orthologs roughly defines the 'core proteome' likely to be encoded by most contemporary eukaryotes. Because basic cellular processes are largely conserved across eukaryotes, we assume that the relative abundances of proteins in these processes should be conserved as well, roughly maintained by selection. If this were indeed the case, it would represent a unique opportunity to externally validate any protein quantification technique: the best technique would be the one that results in the highest correlation of observed abundances across the five organisms. For spectral counting algorithms, comparisons among such distant organisms (e.g.: yeast *vs*. human) provide another important advantage: the proteins differ sufficiently in sequence, meaning that physical or technical biases per peptide are mostly averaged out.

We find that our initial assumption indeed holds: proteins in the conserved core are significantly correlated in their abundance, and the residual variance is not simply random but instead informative of evolutionary and functional constraints, as described below.

Materials and Methods

Data sources

We imported MS/MS protein identifications from the PeptideAtlas database [27], namely from the builds dated March and April 2009 for data from yeast and human, respectively. For *C. elegans*, *D. melanogaster* and *A. thaliana*, we employed MS/MS data directly from dedicated, genome-wide projects published previously [18-20] (Suppl. Fig. 1; in the case of *C. elegans*, our data includes a number of additional, more recent samples, extending the spectral counts by roughly 50% over the published counts). The data that we assembled cover a variety of tissues, cell-lines, environmental conditions and/or developmental stages, and diverse protocols for biochemical fractionation had been applied. For all five organisms, the Trans-Proteomic Pipeline had originally been used for protein identification (http://tools.proteomecenter.org/TPP.php). In this pipeline, and in PeptideAtlas, a uniform cutoff for the reliability of peptide identifications is used (i.e. PeptideProphet score >= 0.9).

Protein abundance quantification

We used a simple spectral counting algorithm employed previously [20], to estimate the protein abundance from the frequencies of observed peptide spectra. For any theoretical tryptic peptide within a protein, we first estimated its likelihood of being successfully identified from MS/MS data, based on its length – a dependency that can be learned from the data and that is roughly the same for all five organisms studied here [Suppl. Fig. 2]. The actual spectral counts were then weighted by this length-based detectability factor (see below). We did not account for more elaborate physical properties of the peptides [24, 25], for three reasons: a) our analysis contains data from several laboratories, and not all of these have used the exact same setup for sample processing and mass spectrometry acquisition, b) in our hands, length is the most important determinant of peptide detectability, and c) sequence-based algorithms could potentially be subtly over-

trained, which could result in spuriously high abundance correlations for orthologs that have high sequence conservation. Using our simplified approach, the latter problem did not occur: abundance correlations were not higher for proteins of high sequence conservation (i.e., abundance-corrected variance and sequence conservation were not correlated; data not shown). Similarly, we chose not to exclude peptide observations from certain specific experimental setups (such as ICAT, which enriches for Cys-containing peptides). We did test the removal of all Cys-containing peptides, but this resulted in lower correlations against external references, meaning that Cys-containing peptides currently deliver more signal than noise.

Still, there remains the possibility that a given pair of proteins appear to have a similar abundance simply because their constituent tryptic peptides have the same intrinsic 'MS-detectability' (i.e., beyond the peptide-length effect for which we already correct). To test for this, we artificially suppressed equivalent peptides, by only using alternating sections of the aligned proteins [Suppl. Fig. 3]. As a control, we downsampled the data by the same amount, but this time always considered equivalent peptide positions only. This test does reveal a small, residual effect of shared peptide detectability: for example in the case of the human/yeast comparison, using alternating peptides results in a correlation of R_S=0.494, but using equivalent peptides gives a slightly higher correlation of R_S=0.518. However, this residual effect is quite small, and it has no impact on our further conclusions. Additional support for the validity of our abundance estimates lies in their correlation with independent, known surrogates for protein abundance [Suppl. Fig. 4]: our data show a correlation of R_S=0.65 with the 'codon adaptation index' (CAI) in yeast, and an inverse correlation (R_S= -0.27) with protein length.

We describe all individual protein abundances in 'parts per million', relative to the molecule counts of all other proteins in the detected proteome (or, alternatively, with reference to the core proteome only). We did not consider splice-variants separately, i.e. our measurements are 'locus-based': all splice-variants encoded by a locus are aggregated and contribute jointly to a single abundance value for this locus. The actual abundance values were computed as follows: we counted how often any of its amino acids had been identified in a protein, divided by the total number of amino acids in the protein sequence (the latter being length-corrected as described above). An additional length restriction

limiting peptides to within ≥ 7 and ≤ 40 amino acids (modified from [24]) was applied, and final counts were normalized and expressed as parts per million.

$$a = \frac{\sum_i number(p_i) \cdot length(p_i)}{\sum_j length(q_j) \cdot f(q_j)}$$

a = protein abundance
p = identified peptides
q = tryptic peptides (*in silico* digest)
f(q) = peptide length correction factor

It should be stressed that the abundance estimates we thus derive are still subject to considerable error. For example, upon randomly splitting the yeast peptide data in half, the corresponding abundance estimates only correlate to each other with R_S=0.92. From this, we infer a conservative average error of around two-fold for the individual estimates [Suppl. Fig. 5]. This error is even larger when splitting the data not randomly, but along different experimental samples [Suppl. Fig. 5], showing that there are noticeable systematic differences among the various procedures and platforms used.

Orthologs

We constructed a dedicated set of orthologous groups covering the five organisms we studied here. For this, we imported protein sequences from version 8.0 of the STRING database [28], which contains a complete, pre-computed set of all-against-all BLAST homology relations for these sequences. From the homology relations, we computed orthologous groups essentially as originally proposed by Tatusov et al [29], in an implementation written for the eggNOG database [30]. Briefly, the groups are constructed by joining 'triangles' of reciprocal-best-match relations, each involving three species. Triangles are joined when they share one edge, whereby the highest-scoring triangles are joined first. Prior to triangle formation, we search for proteins that are more closely related to each other within an organism than to any of the proteins in the other four organisms ('inparalogs'). These inparalogs are grouped and represented by their highest-scoring member in the subsequent triangle searches. After triangle joining, all pair-wise alignments are tested to verify that the proteins in a group can all be aligned to each other, in a way that defines at least one common sequence segment.

Expression variance

The 1581 orthologous groups present in all five organisms were used to define the core proteome, and for 1172 of these we were able to infer all five abundance estimates (one for each organism). In the case of organisms having more than one protein in an orthologous group, the abundances of these inparalogs in the organism were added up. The observed normalized variance per group ("CV", coefficient of variation) was inversely related to protein abundance (Figure 3; $p<10^{-15}$). To account for this effect and to obtain an 'intrinsic' measure of variance that could differentiate the variance of proteins independent of their abundance, we ranked CV values by abundance and computed a running median (Fig. 3). Then, for each protein family, the difference between the observed variance and the median variance at that particular abundance range (the "DM" value: 'distance to median') was introduced as a measure for the 'intrinsic variance' (similar to the procedure in [4]). We express DM values in normalized form (i.e., in percent of the median value), with negative values indicating less variance than expected, and positive values indicating more variance than expected.

Functional annotations

Since yeast is arguably among the best-annotated organisms (in terms of molecular biology details), we categorize orthologous groups of proteins based on the annotation of their yeast protein member(s). Gene ontology annotations of yeast proteins were imported from SGD [31], and an updated catalog of protein complexes was imported from CYC2008 [32].

Statistical tests

Unless otherwise noted, two-sided Kolmogorov-Smirnov tests were used to gauge the significance of observed differences in distributions. To correct for multiple testing, p-values were adjusted according to Benjamini and Hochberg [33]. All abundance correlations are rank-based (Spearman correlations), and p-values for these correlations were computed by the "AS 89" algorithm as implemented in the R software package.

Results

In order to characterize the eukaryotic core proteome and its compositional stability, we first defined a set of protein families universally encoded in the five organisms we studied

(*S. cerevisiae*, *A. thaliana*, *H. sapiens*, *D. melanogaster* and *C. elegans*). We found 1581 such protein families, usually represented by exactly one protein-coding locus in a given organism (584 families contained a single locus in all five genomes; the average representation of all 1581 groups is 1.41 loci per organism). This set of proteins likely represents a nearly universal, ancient core of eukaryotic proteins – comprised mostly of proteins with information processing functions, metabolic functions, and cellular maintenance functions.

Independent of the orthology detection, we also estimated the abundance of all detectable proteins in these five organisms, based on about six million peptide spectra available from public databases (Figure 1 and Suppl. Fig 1; see also Methods). This resulted in protein abundance values extending over more than four orders of magnitude, from less then 1 ppm (parts per million) to more than 10'000 ppm. Of the 1581 protein families having representatives in the five organisms, 1172 received an abundance estimate in all five organisms – this is the set of protein families we used for all evolutionary analyses described below.

In the case of yeast, four independent experimental datasets on protein abundance are available [4, 12, 24, 34], two of which are not based on mass spectrometry and can therefore be used for validation. We find that those two latter sets correlate reasonably well with our estimates based on spectral counting (R_S = 0.65 and R_S = 0.58, respectively; [Suppl. Fig. 6]). However, why are these correlations not higher? They could in principle be limited by errors on both sides – errors within the experimental measurements and/or errors within our spectral counting technique. To investigate this, we used transcript abundances as an independent 'arbiter': we compared the data sets against absolute transcript abundances in yeast (and also against transcript abundances in *C. elegans,* the latter being an arbiter at a much larger evolutionary distance). Despite the limited overall correlation between transcript and protein abundances [35, 36], transcripts can serve as 'arbiters' because they share little technical biases with proteins: a gene that encodes a problematic protein (e.g. a trans-membrane protein) may still encode a transcript that can be reliably measured, and *vice versa* (e.g. a transcript with problematic secondary structure may still encode a protein that is well-suited for mass spectrometry). Remarkably, in all tests our spectral-counting-derived protein abundances showed the

best correlations against the transcript levels [Suppl. Fig. 7]. Together with the fact that the spectral counting data have a much higher coverage than the biochemical experimental data, this suggests that they are indeed among the best quantitative protein abundance data available in any eukaryote to date (rivaled, in yeast only, by recent SILAC data [34], which correlates slightly lower but has better coverage). Globally, we covered about 56% of all predicted proteins encoded in the five genomes (ranging from only 48% in plants to about 60% in fly). Note that, within yeast, the best correlation is seen between the two MS-based datasets (our spectral counting *vs.* the SILAC data; R_S=0.70; Suppl. Fig. 6). This is notable because distinct measurement strategies have been used for the two sets (area under the peak, and spectral counting, respectively). Still, this correlation is far from perfect, and is a reminder of the relatively high level of noise in any genome-wide protein abundance quantification (see Methods section for noise quantification).

By mapping our abundance values onto the orthology information, we then obtained five independent snapshots of essentially the same core proteome, separated not only by hundreds of million years of independent evolution, but also by differences in sampling material and procedures (e.g. various growth conditions, differences in handling, tissue selection and sample preparation). Because of these evolutionary and environmental/procedural differences, the five snapshots should represent a good view of the actual compositional variance of the core proteome. It should be noted that the core proteome represents only a rather small part of the full proteomes (covering 20.7% of all proteins in yeast, but only 7.6% of the proteins in human). In addition, we are effectively 'averaging' over various conditions and stages, meaning that we cannot study the varied expression of individual proteins in response to external stimuli. Nevertheless, the averaged core proteome enables comparisons between the five organisms on an equal footing. In addition, it consists mainly of 'house-keeping' genes whose tissue-to-tissue expression variability is limited, and it encompasses many essential functions such as information processing, cellular structure, metabolism and signal transduction.

To analyze the results, we first computed the rank abundance correlations between the various core proteomes, which we interpret as a lower limit of their quantitative conservation. As reported earlier for the case of *C. elegans* vs. *D. melanogaster* [20], we generally found these correlations to be surprisingly high, ranging from R_S=0.57 for the

comparison yeast vs. plant, to R_S=0.85 for the comparison nematode vs. fruit fly (Figure 2). Most likely, the underlying biological correlations are in reality even higher, since spectral counting is presumably limited by the numerical noise that is associated with under-sampling. If limited sampling is indeed an issue, then we should be able to increase the correlations by including more data – and this is what we observe: we can improve the correlation between the non-MS-derived, independently measured protein abundances in yeast and our MS-derived abundances by averaging the MS data from an increasing number of other species (Figure 2d). The correlations across large evolutionary distances were also improved by aggregation, for example by grouping the three animals and comparing them to a grouping of yeast and arabidopsis (R_S=0.80; this compares to R_S in the range of 0.64 to 0.70 for the individual comparisons at this distance; note that yeast and arabidopsis are not a meaningful phylogenetic grouping but were merely grouped because both are at large distances from animals). This latter comparison of course does not only increase sampling depth, but also smoothes out individual differences among the organisms.

Given the remarkable evolutionary conservation of protein abundances, we next tested whether groups of functionally linked proteins would be expressed at distinct, typical abundances. For this, we grouped proteins according to their membership in stable yeast protein complexes [32]. Indeed, we observe that members of the same protein complex tend to have similar abundances [Suppl. Fig. 8]. This observation suggests that we can expect to further decrease numerical noise by grouping proteins by their membership in protein complexes – this is what we find, resulting for example in an abundance correlation of R_S=0.92 between entire protein complexes in the worm on the one hand and the fly on the other hand [Suppl. Fig. 9]. Interestingly, we could not further increase the correlation across the animal *vs.* fungi/plant grouping, suggesting that our observed correlation of R_S=0.80 is already close to the actual biological correlation at this large distance.

Next, we studied the abundance *variance* across species, for each individual protein family (Figure 3; see Methods). We first observed that the variance scales inversely with protein abundance, as has been observed for cell-to-cell expression variance as well [4]. In our case, it is likely that this observation is at least partly due to technical limitations –

instrument error and spectral counting inaccuracies are presumably more prevalent for proteins of low abundance. Therefore, we normalized the variance data for protein abundance and obtained an 'intrinsic' variance measure termed DM (distance from a running median of variances), in accordance with [4]. These variance values are still very noisy in themselves (given that each variance calculation is based on only five data points), but they provide – for the first time – an objective view on the intrinsic quantitative conservation of the eukaryotic core proteome. In order to be biologically meaningful, these variances should correlate with externally described functional properties of the proteins, which is what we investigated next.

First, we grouped proteins into broad functional categories as described in the Gene Ontology database (GOslim) [37]. We observed that proteins in the biological process category 'Transport' show unusually high variance, whereas proteins in the categories 'Translation' and in particular also 'Ribosome biogenesis' showed unusually low variance (Figure 3). Similarly, we observed significant differences among cellular localizations and molecular functions, most notably a lower-than-expected variance of proteins in the categories 'Transcriptional regulator' and 'Translational regulator'. We also observed a weak but significant correlation of our variance with cell-to-cell expression noise (R_S=0.11, p=0.003; [Suppl. Fig. 10]). Next, we turned to essentiality data in yeast [38]; here again, we observed differences in variance: the variance is significantly lower for essential genes as compared to non-essential genes (p=0.0001; Figure 4).

For functional groupings, the combination of low variance across species and good intra-group agreement of abundance values across genes suggest that global stoichiometries between functional groups might be relatively constant across organisms. As a case in point, we computed for all five organisms the relative stoichiometries of the categories 'Translation' *vs.* 'Ribosome biogenesis'. We find that this ratio is about 20:1, and remarkably, it is relatively stable across all five organisms (ranging from 13:1 in human and arabidopsis to 23:1 in fruit fly).

The last aspect we addressed was gene duplicability – many historical instances of gene duplications presumably affected the abundance of the encoded proteins. Thus, protein families whose abundance variations are more tightly controlled might have experienced a stronger purifying selection against fixation of a duplicate copy. To test this, we

classified orthologous groups according to the number of paralogs they contain (this roughly reflects the frequency of retained gene duplication events in the five lineages). We observe that the groups with the lowest numbers of paralogs indeed showed the lowest variance ($p=0.0013$; Figure 4), suggesting that the need to maintain a given protein abundance level may at least partly be responsible for restrictions on gene duplicability.

Discussion

Biological variation of protein abundances has been studied extensively, but so far only at the level of individual cells [4, 5, 39, 40], or across closely related species [41]. In contrast, here we study the extent to which protein expression levels can vary over large evolutionary distances, while we average over distinct tissues and/or environmental conditions. Thus, we are essentially probing the upper limit of the quantitative compositional flexibility of the 'core' proteome in the course of evolution.

Many parts of the proteome are of course heavily regulated, and will change their expression levels over orders of magnitude in response to external stimuli or developmental cues. However, what we study here is a unique and stable subset of the proteome – the 'core' as defined by ubiquity across eukaryotic kingdoms. A significant part of this core is likely expressed in all cells and under all conditions. Indeed, we observe that of the 1581 protein families we studied, 75% can be detected by mass spectrometry in all five organisms (this figure increases to 92% when lowering the requirement to four organisms). Proteins in the core appear generally better conserved and/or better suited for quantification: within yeast, for example, the two MS-based datasets (SILAC and spectral counting) correlate better for core-proteins ($R_S=0.75$) than for non-core proteins ($R_S=0.62$), the former correlation being higher than any inter-organism correlation involving yeast.

For this subset – the 'eukaryotic core proteome' – we have now established a reference of typically observed protein abundance ranges, and this can serve as a baseline against which to compare cellular protein abundance states in the future. To us, the observed evolutionary conservation suggests purifying selection: proteins would be roughly maintained at optimal abundance levels – whereby functional requirements would limit the amount of under-production, and both toxicity and the metabolic cost of protein

production would limit over-production. The metabolic cost of protein production can be visible to selection in eukaryotes, at least in species with large effective population sizes [42]. While complex eukaryotes such as humans may be relatively indifferent to the metabolic cost of protein production, at least for lowly expressed individual proteins, the core proteome examined here consists largely of highly expressed proteins, for which total production cost may be appreciable. That expression levels are to some extent maintained by purifying selection has already been proposed based on transcript abundance data, although the extent of selection at the transcript level remains controversial [7, 8, 43-46].

Invoking selection in general is also compatible with our observation that essential genes are better maintained at their typical abundance level than non-essential genes, presumably because of both, abundance mismatches having greater fitness consequences for these genes, and essential proteins being often highly expressed [47, 48] and thus imposing a larger metabolic burden on the cell. Furthermore, as the overall cost of protein production is proportionally higher for more abundant proteins, a role of production costs in limiting abundance variation is also consistent with our finding of a lower coefficient of variation (CV) for highly abundant proteins. Most striking, however, is the implicit conclusion that core protein stoichiometries actually do have an optimum that applies across eukaryotic kingdoms. This suggests a detailed quantitative conservation of the core cellular processes, independent of vastly different physiologies and cellular organizations. Thus, to modify a quote of Jacques Monod, “what is true for yeast is true for the elephant”, not just in principle but also in rather surprising quantitative detail.

Interestingly, we have not observed a correlation between our abundance variance and protein evolutionary rate (data not shown). This appeared puzzling at first, as proteins with more conserved expression levels might intuitively also be functionally more constrained, and thus perhaps also in their abundance variance. However, evolutionary rate at the sequence level appears to be largely dominated by constraints from protein folding and not necessarily by functional constraints [49, 50]. In contrast, our variance would mainly be controlled by constraints at the level of gene regulation, for example by keeping production levels and turnover rates of both transcript and protein relatively constant. From the outset, protein evolutionary rate could have been a potential

confounding factor of our analysis, as it is known to be correlated with abundance itself, and to some extent also with essentiality (which, in turn, is correlated with gene duplicability). However, when we performed a step-wise multiple linear regression, testing these and other variables, they were not detected as confounding factors [Suppl. Fig. 11]. For abundance variance, the most important predictor remains abundance itself, as shown in Figure 3a. After correcting for this, the number of duplicated genes follows, and then essentiality and cell-to-cell noise [Figure 4; Suppl. Fig. 11].

Despite the strong overall correlations of protein abundances, we did of course observe differences in the abundance of individual proteins across species. Do these differences reflect adaptations, short-term gene regulation due to distinct stimuli, or are they the consequences of genetic drift? All else being equal, drift in abundance may be expected to be stronger for proteins that can tolerate higher levels of expression noise (cell-to-cell variation at given environmental conditions). That our variance at evolutionary timescales correlates only very weakly with cell-to-cell protein expression noise therefore suggests that many of our observed differences are perhaps not due to drift, but may reflect regulation or adaptations to the substantial physiological differences among the eukaryotes studied here.

Finally, we observed that gene families with more duplicates tend to show higher variance in protein expression levels. This argues against a strict model of sub-functionalisation, in which the functions of the ancestral protein would be simply divided up between the duplicate copies; in this case, their total abundance should remain identical and our variance would not be affected – since we add up the contributions of all paralogs in a family. On the other hand, total abundance levels do not grow linearly with the number of paralogs either [20], and hence duplications most likely fix due to a mixture of sub- and neo-functionalisation.

In summary, our analysis provides the first quantitative overview of the core eukaryotic proteome, and helps to further establish spectral counting as a semi-quantitative measure (provided that a good sampling depth can be achieved). It will be intriguing to extend the analysis to more organisms and to deeper spectral counts, in order to achieve a more fine-grained view on protein abundance stability, on purifying selection, and perhaps also on

ways to objectively separate the 'constitutive' from the 'regulated' parts of any given proteome.

Figure Legends

Figure 1: Tryptic peptide observations on aligned orthologs in five species
The yeast protein UBC1 (a Ubiquitin-conjugating enzyme), and four of its orthologs in multicellular eukaryotes are shown aligned; the predicted tryptic cleavage sites are marked in red. Peptide observations, from a collection of published shotgun proteomics experiments, are pooled and shown as horizontal lines above the protein sequences; each line represents one peptide observation. Inset: magnified section showing details for the yeast/plant alignment; conserved residues are indicated and the plant peptides are shown below the sequence for clarity. Lower right: abundance computation for individual proteins. The observed peptides are normalized by the theoretically expected tryptic peptides in a given protein; the latter are corrected for 'observability' based on their length.

Figure 2: Protein abundance correlations
a) Comparison of the inferred abundances of orthologous proteins in two organisms (yeast vs. human). **b)** Increased abundance correlation upon data aggregation: comparison of the average abundances of yeast and plant, vs. the average abundance of the three animals. **c)** Spearman rank correlations for all pairwise comparisons. **d)** Data aggregation across organisms also improves the correlation against non-MS derived, independently measured protein abundances in yeast (in this case, the yeast MS data were left out). All abundances in this figure are indicated relative to the core proteome only (i.e. non-conserved proteins are not contributing to the total when computing the abundances; the indicated values denote relative molecule counts in parts-per-million [ppm]).

Figure 3: Variance of protein abundance

a) Coefficient of variation (i.e., variance divided by the mean) across five organisms, plotted against median abundance (each dot represent an orthologous group of proteins present in five organisms). A running median is shown, with a window size of 200 and truncated window-sizes at both ends. **b)** Two exemplary proteins with similar median abundance, but large differences in their coefficients of variation are shown. GPA2 is a G-protein subunit involved in glucose sensing, whereas UBC1 is a Ubiquitin-conjugating enzyme that helps degrade short-lived or abnormal proteins. GPA2 is presumably strongly regulated; hence the extreme differences in observed abundances. **c)** Same plot as in a), with three high-level functional categories marked in color. **d)** Normalization against the abundance dependency: In this plot, variance is now expressed as 'distance to the running median' (DM), indicated in percent of the median. **e)** Distributions of DM-values, for all proteins (grey) and for three functional categories shown in color (all three differ significantly from the background distribution; $p = 0.002$, $p=0.003$ and $p<1e-05$, respectively).

Figure 4: Abundance variance and evolutionary signals

a) The abundance variance of an orthologous group increases with the number of inparalogs in that group. **b)** Essential genes have a lower abundance variance than non-essential genes. Both plots are based on abundance-corrected variance measures (DM), as in Figure 3.

Figure S1: Overview of the data sources

In total, about 6 million mapped peptide spectra were considered for this study. The table provides an overview of the number and size of experimental samples, and gives an indication on the experimental platforms used (to the extent that this information is annotated in PeptideAtlas).

Figure S2: Length-dependency of peptide detections in MS/MS experiments

For all five organisms, the predicted distribution of peptides derived from a hypothetical, complete tryptic digest of the encoded proteome, is contrasted to the actual distribution

of observed and successfully identified peptides. Both distributions are characteristic, and are largely reproducible between the five organisms.

Figure S3: Testing for correlated detectability of homologous peptides

a) To artificially exclude 'equivalent' peptides from the abundance computations, alternating sections of aligned proteins are blocked from consideration. Conserved tryptic cleavage sites serve as 'anchors', and the procedure results in a down-sampling of the data by roughly 50%. **b)** Control experiment: here, equivalent peptides are preferred for the abundance computations, and the amount of down-sampling is comparable.

Figure S4: Testing the MS-based abundance against independent variables

a) Protein abundance, as estimated based on spectral counting, is plotted against the 'codon adaptation index' (CAI) [51]. **b)** Spectral-counting based abundance is plotted against protein length.

Figure S5: Error-modeling of the abundance estimate, based on self-correlation of yeast data

a) All yeast peptide identifications were randomly divided into two groups; the resulting protein abundance estimates are shown plotted against each other. **b)** Plotting one of the two sets against itself results in a perfect correlation. **c)** Same plot as in b), except that on both axes, a controlled amount of noise has been added (that noise has a linear dependency on log-abundance). Vertical lines denote two arbitrary positions where the level of noise is indicated. **d)** through **f)**: the same test was executed, but the separation of the yeast data into two groups was now along annotated experiments: experiments were randomly assigned to two groups such that the final groups were roughly of the same size.

Figure S6: Various available protein abundance datasets in yeast, and how they compare

a) Table with pair-wise Spearman correlation coefficients. All datasets were reduced to a common set of proteins before the comparisons. **b)** Plot showing the best-correlated pair: our abundance estimates vs. the genome-wide SILAC measurements of ref [34].

Figure S7: Validation and ranking of proteome quantification datasets

a) Proteome-wide quantification data on absolute proteome abundance are compared against absolute transcript abundance levels in yeast [52]. Transcriptomics data are from a completely unrelated set of technologies, and while the biological correlation is known to be imperfect, transcript abundances can act as an unbiased 'arbiter': the better the quantification of the proteome, the better it should correlate with the transcriptome. **b)** All four genome-wide protein abundance datasets are compared to the same transcriptomics experiment, reduced to a set of proteins covered in each of them. **c)** The comparison against the *C. elegans* transcriptome is shown. Here, correlation coefficients are generally lower (due to the large evolutionary distance), but there should be less possibility for shared biases (growth conditions, etc). Note that the relative ranking is identical.

Figure S8: Proteins in the same complex tend to have similar abundances

Pairs of proteins from the same protein complex (left), or an equivalent number of pairs with random complex membership (right) are shown, and the distribution of abundance differences for the pairs is plotted (in log-space). Note that proteins annotated in the same complex tend to have much lower abundance differences. Protein complex membership was parsed from the CYC2008 collection of yeast complexes [32].

Figure S9: Aggregated protein abundances per complex: high inter-species correlation

Yeast proteins were grouped into protein complexes as described in [32], and this mapping was projected onto the fly and worm proteomes using our orthologous groups (this assumes that protein complexes are generally well-conserved regarding their constituent proteins [53]). Then, all protein abundances in a complex were added up. Each dot represents one distinct protein complex.

Figure S10: Cell-to-cell protein expression noise correlates only weakly with expression variance at evolutionary time-scales.

Intrinsic cell-to-cell abundance variance of yeast proteins ('dm', as defined in [4]), are shown compared to abundance-corrected variance at evolutionary distances ('dm', as

defined in our study, based on orthologs in five organism). Each dot represents one orthologous group containing a yeast protein and at least one protein in the four other organisms.

Figure S11: Step-wise multiple regression in search of factors that can partially explain the protein abundance variance across organisms.

In Figure 4 of the main text, we report that the variance of protein abundance across organisms is correlated with the number of paralogs, and with essentiality as measured in yeast. To test whether these two correlations are independent, and/or whether there exist any confounding variables explaining either of these two effects, we performed a step-wise multiple regression analysis based on generalized linear models (as implemented in the 'glm' function in R). Before analysis, all variables were converted to ranks, so that the correlations shown are Spearman correlations throughout. **a)** individual correlations of various variables with the normalized and corrected protein abundance variance ('dm_percent'). **b)** all variables that showed significant correlations above were subsequently joined into a generalized linear model. Within that model, variables that had no significant contributions were successively removed. The three remaining variables are shown – note that there were no statistically significant interactions among these (i.e., $p > 0.05$ in all pair-wise combinations).

Table S1: Normalized abundance data for the core proteome

Each data line describes a single orthologous group of proteins, which is represented in all five organisms studied. The respective protein abundances in each organism, as well as the constituent genes, are indicated.

Table S2: The eukaryotic core proteome sorted according to its expression variance

Each data line describes a single orthologous group of proteins. The list is sorted according to the abundance-corrected variance between the organisms ('percent dm').

Acknowledgements

This work has been supported by the University of Zurich through its Research Priority Program "Systems Biology and Functional Genomics", by the Gerbert Rüf Foundation, by the Ernst Hadorn Foundation and by the Swiss National Science Foundation.

Conflict of interest statement

The authors declare that they have no conflicts of interest, neither financial nor otherwise.

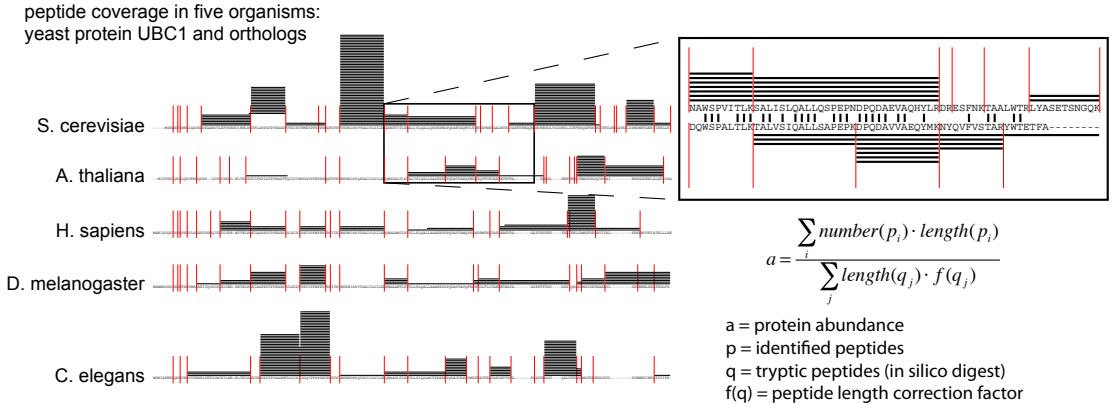
References

[1] Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E.*, et al.*, Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007, *315*, 848-853.
[2] Henrichsen, C. N., Chaignat, E., Reymond, A., Copy number variants, diseases and gene expression. *Hum Mol Genet* 2009, *18*, R1-8.
[3] Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M.*, et al.*, Noise in protein expression scales with natural protein abundance. *Nat Genet* 2006, *38*, 636-643.
[4] Newman, J. R., Ghaemmaghami, S., Ihmels, J., Breslow, D. K.*, et al.*, Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* 2006, *441*, 840-846.
[5] Cohen, A. A., Geva-Zatorsky, N., Eden, E., Frenkel-Morgenstern, M.*, et al.*, Dynamic proteomics of individual cancer cells in response to a drug. *Science* 2008, *322*, 1511-1516.
[6] Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I.*, et al.*, A neutral model of transcriptome evolution. *PLoS Biol* 2004, *2*, E132.
[7] Bedford, T., Hartl, D. L., Optimization of gene expression by natural selection. *Proc Natl Acad Sci U S A* 2009, *106*, 1133-1138.
[8] Yanai, I., Graur, D., Ophir, R., Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS* 2004, *8*, 15-24.
[9] Fay, J. C., Wittkopp, P. J., Evaluating the role of natural selection in the evolution of gene regulation. *Heredity* 2008, *100*, 191-199.
[10] Dekel, E., Alon, U., Optimality and evolutionary tuning of the expression level of a protein. *Nature* 2005, *436*, 588-592.
[11] Milo, R., Jorgensen, P., Springer, M., *http://bionumbers.org*.
[12] Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W.*, et al.*, Global analysis of protein expression in yeast. *Nature* 2003, *425*, 737-741.
[13] Ong, S. E., Mann, M., Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* 2005, *1*, 252-262.

[14] Pan, S., Aebersold, R., Chen, R., Rush, J., *et al.*, Mass spectrometry based targeted protein quantification: methods and applications. *J Proteome Res* 2009, *8*, 787-797.
[15] Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* 2007, *389*, 1017-1031.
[16] Malmstrom, J., Beck, M., Schmidt, A., Lange, V., *et al.*, Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* 2009, *460*, 762-765.
[17] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008, *26*, 1367-1372.
[18] Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H., *et al.*, A high-quality catalog of the Drosophila melanogaster proteome. *Nat Biotechnol* 2007, *25*, 576-583.
[19] Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., *et al.*, Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* 2008, *320*, 938-941.
[20] Schrimpf, S. P., Weiss, M., Reiter, L., Ahrens, C. H., *et al.*, Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biol* 2009, *7*, e48.
[21] Gao, J., Opiteck, G. J., Friedrichs, M. S., Dongre, A. R., Hefta, S. A., Changes in the protein expression of yeast as a function of carbon source. *J Proteome Res* 2003, *2*, 643-649.
[22] Liu, H., Sadygov, R. G., Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 2004, *76*, 4193-4201.
[23] Ishihama, Y., Oda, Y., Tabata, T., Sato, T., *et al.*, Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* 2005, *4*, 1265-1272.
[24] Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E. M., Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 2007, *25*, 117-124.
[25] Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., *et al.*, Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* 2007, *25*, 125-131.
[26] Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., *et al.*, Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J Proteome Res* 2006, *5*, 2339-2347.
[27] Deutsch, E. W., Lam, H., Aebersold, R., PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* 2008, *9*, 429-434.
[28] Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., *et al.*, STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 2009, *37*, D412-416.
[29] Tatusov, R. L., Koonin, E. V., Lipman, D. J., A genomic perspective on protein families. *Science* 1997, *278*, 631-637.
[30] Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., *et al.*, eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* 2008, *36*, D250-254.
[31] Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R., *et al.*, Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res* 2008, *36*, D577-581.

[32] Pu, S., Wong, J., Turner, B., Cho, E., Wodak, S. J., Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res* 2009, *37*, 825-831.
[33] Benjamini, Y., Hochberg, Y., Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B* 1995, *57*, 289-300.
[34] de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L.*, et al.*, Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 2008, *455*, 1251-1254.
[35] Gygi, S. P., Rochon, Y., Franza, B. R., Aebersold, R., Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 1999, *19*, 1720-1730.
[36] Fu, X., Fu, N., Guo, S., Yan, Z.*, et al.*, Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* 2009, *10*, 161.
[37] Harris, M. A., Clark, J., Ireland, A., Lomax, J.*, et al.*, The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004, *32*, D258-261.
[38] Giaever, G., Chu, A. M., Ni, L., Connelly, C.*, et al.*, Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 2002, *418*, 387-391.
[39] Raser, J. M., O'Shea, E. K., Noise in gene expression: origins, consequences, and control. *Science* 2005, *309*, 2010-2013.
[40] Becskei, A., Kaufmann, B. B., van Oudenaarden, A., Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat Genet* 2005, *37*, 937-944.
[41] Fu, N., Drinnenberg, I., Kelso, J., Wu, J. R.*, et al.*, Comparison of protein and mRNA expression evolution in humans and chimpanzees. *PLoS ONE* 2007, *2*, e216.
[42] Bragg, J. G., Wagner, A., Protein material costs: single atoms can make an evolutionary difference. *Trends Genet* 2009, *25*, 5-8.
[43] Chan, E. T., Quon, G. T., Chua, G., Babak, T.*, et al.*, Conservation of core gene expression in vertebrate tissues. *J Biol* 2009, *8*, 33.
[44] Yanai, I., Hunter, C. P., Comparison of diverse developmental transcriptomes reveals that co-expression of gene neighbors is not evolutionarily conserved. *Genome Res* 2009.
[45] Xing, Y., Ouyang, Z., Kapur, K., Scott, M. P., Wong, W. H., Assessing the conservation of mammalian gene expression using high-density exon arrays. *Mol Biol Evol* 2007, *24*, 1283-1285.
[46] Gilad, Y., Oshlack, A., Rifkin, S. A., Natural selection on gene expression. *Trends Genet* 2006, *22*, 456-461.
[47] Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M.*, et al.*, Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics* 2008, *9*, 102.
[48] Schmidt, M. W., Houseman, A., Ivanov, A. R., Wolf, D. A., Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe. *Mol Syst Biol* 2007, *3*, 79.
[49] Drummond, D. A., Wilke, C. O., Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 2008, *134*, 341-352.
[50] Powers, E. T., Balch, W. E., Costly mistakes: translational infidelity and protein homeostasis. *Cell* 2008, *134*, 204-206.
[51] Sharp, P. M., Li, W. H., The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 1987, *15*, 1281-1295.

[52] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., *et al.*, The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008, *320*, 1344-1349.
[53] van Dam, T. J., Snel, B., Protein complex evolution does not involve extensive network rewiring. *PLoS Comput Biol* 2008, *4*, e1000132.



$$a = \frac{\sum_i number(p_i) \cdot length(p_i)}{\sum_j length(q_j) \cdot f(q_j)}$$

a = protein abundance
p = identified peptides
q = tryptic peptides (in silico digest)
f(q) = peptide length correction factor

Figure 1

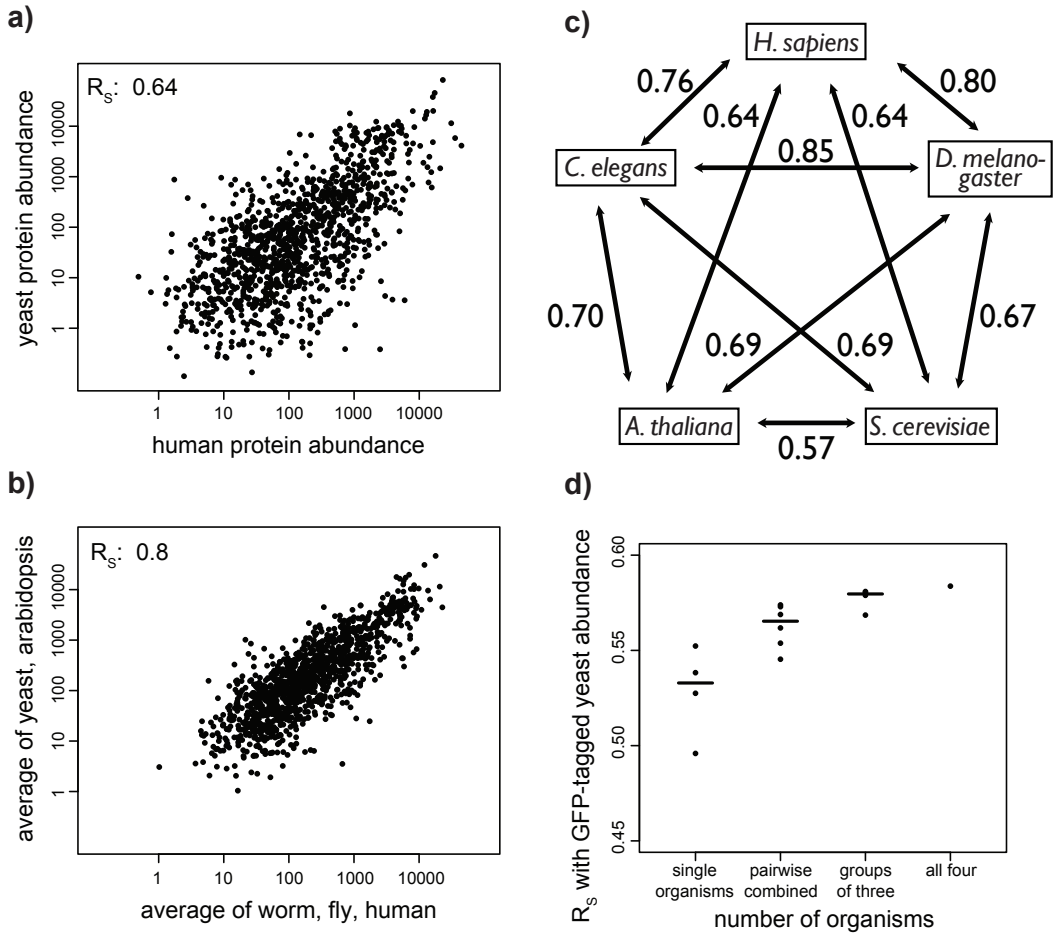

a)
R_S: 0.64
yeast protein abundance
human protein abundance
1
10
100
1000
10000
b)
R_S: 0.8
average of yeast, arabidopsis
average of worm, fly, human
c)
H. sapiens
C. elegans
D. melano-gaster
A. thaliana
S. cerevisiae
0.76
0.80
0.64
0.64
0.85
0.70
0.67
0.69
0.69
0.57
d)
R_S with GFP-tagged yeast abundance
0.45
0.50
0.55
0.60
single organisms
pairwise combined
groups of three
all four
number of organisms

Figure 2

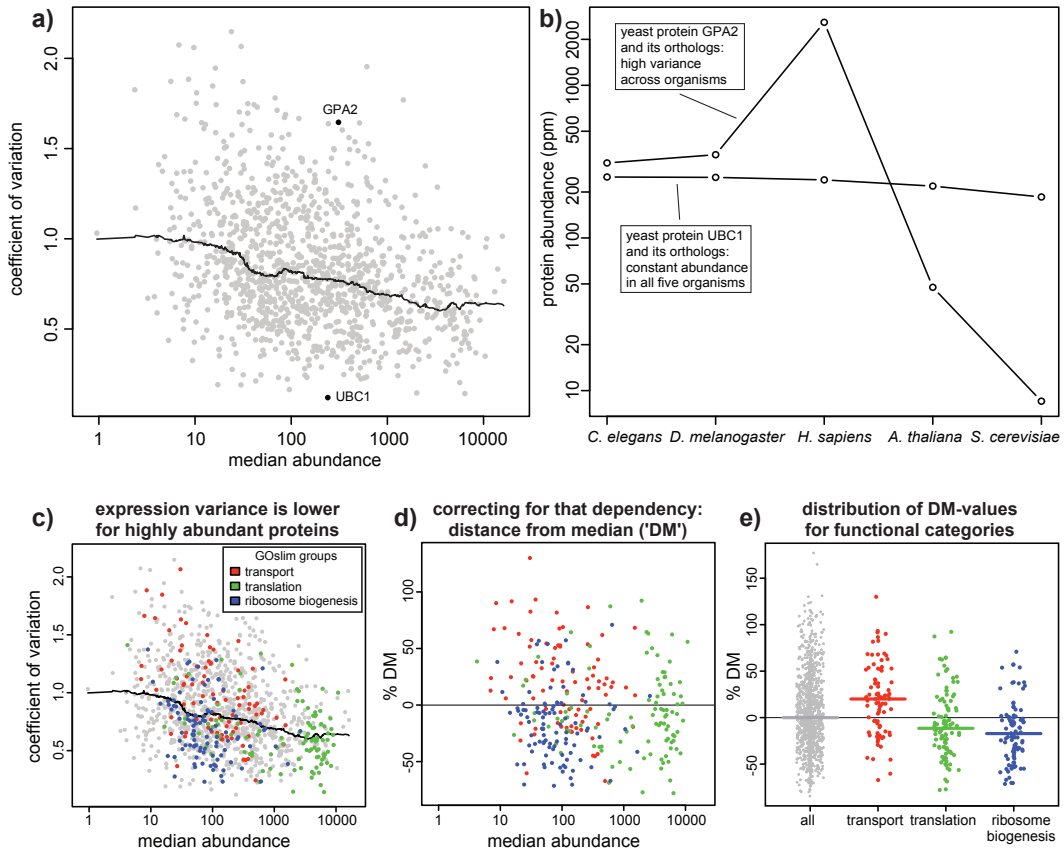

a)
coefficient of variation
median abundance
GPA2
UBC1
b)
protein abundance (ppm)
yeast protein GPA2 and its orthologs: high variance across organisms
yeast protein UBC1 and its orthologs: constant abundance in all five organisms
C. elegans
D. melanogaster
H. sapiens
A. thaliana
S. cerevisiae
c)
expression variance is lower for highly abundant proteins
GOslim groups
transport
translation
ribosome biogenesis
d)
correcting for that dependency: distance from median ('DM')
% DM
e)
distribution of DM-values for functional categories
all
transport
translation
ribosome biogenesis

Figure 3

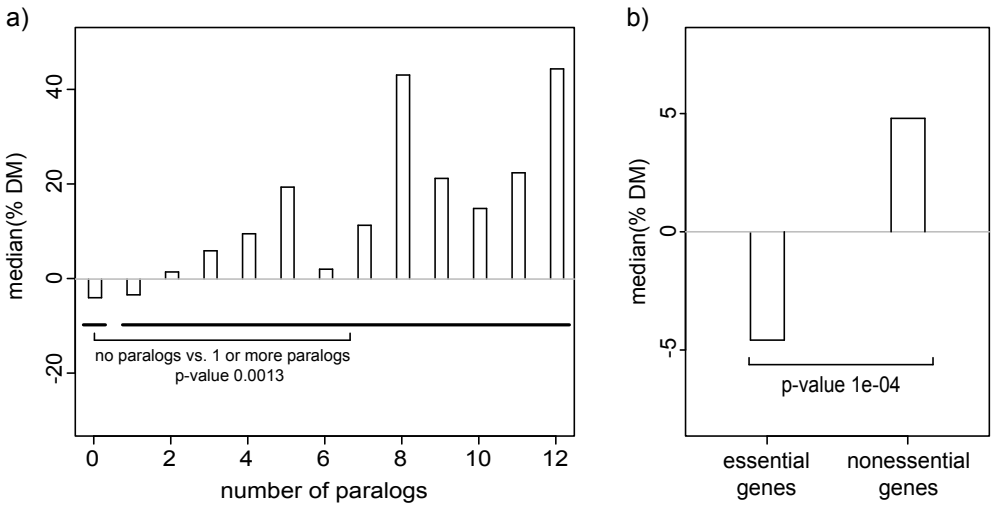

a)
median(% DM)
-20
0
20
40
no paralogs vs. 1 or more paralogs
p-value 0.0013
0
2
4
6
8
10
12
number of paralogs
b)
median(% DM)
-5
0
5
p-value 1e-04
essential genes
nonessential genes

Figure 4

data sets used:

Organism	Name of data set	Number of mapped spectra	Number of unique identified peptides	Number of identified proteins	Number of anno-tated samples/ experiments	Available at
C. elegans	CMOP_Ce	1156565	104309	11728	30	www.peptideatlas.org/repository/
D. melanogaster	*D. melanogaster* PeptideAtlas	498587	72281	8445	44	www.mop.uzh.ch/peptideatlas/
S. cerevisiae	Yeast Build April 2009	1661809	60720	3833	56	www.peptideatlas.org/builds/yeast/
H. sapiens	Human Build March 2009	1926973	83931	12021	123	www.peptideatlas.org/builds/human/
A. thaliana	Pride Accessions 3321-3354	782001	85822	12966	34	www.ebi.ac.uk/pride/

platform types:

S. cerevisiae:
ThermoFinnigan LCQ Classic
ThermoFinnigan LCQ Deca
ThermoFinnigan LCQ Deca XP
ThermoFinnigan LTQ
ThermoFinnigan LTQ FT
ThermoFinnigan LTQ Orbitrap
ThermoScientific LTQ Orbitrap

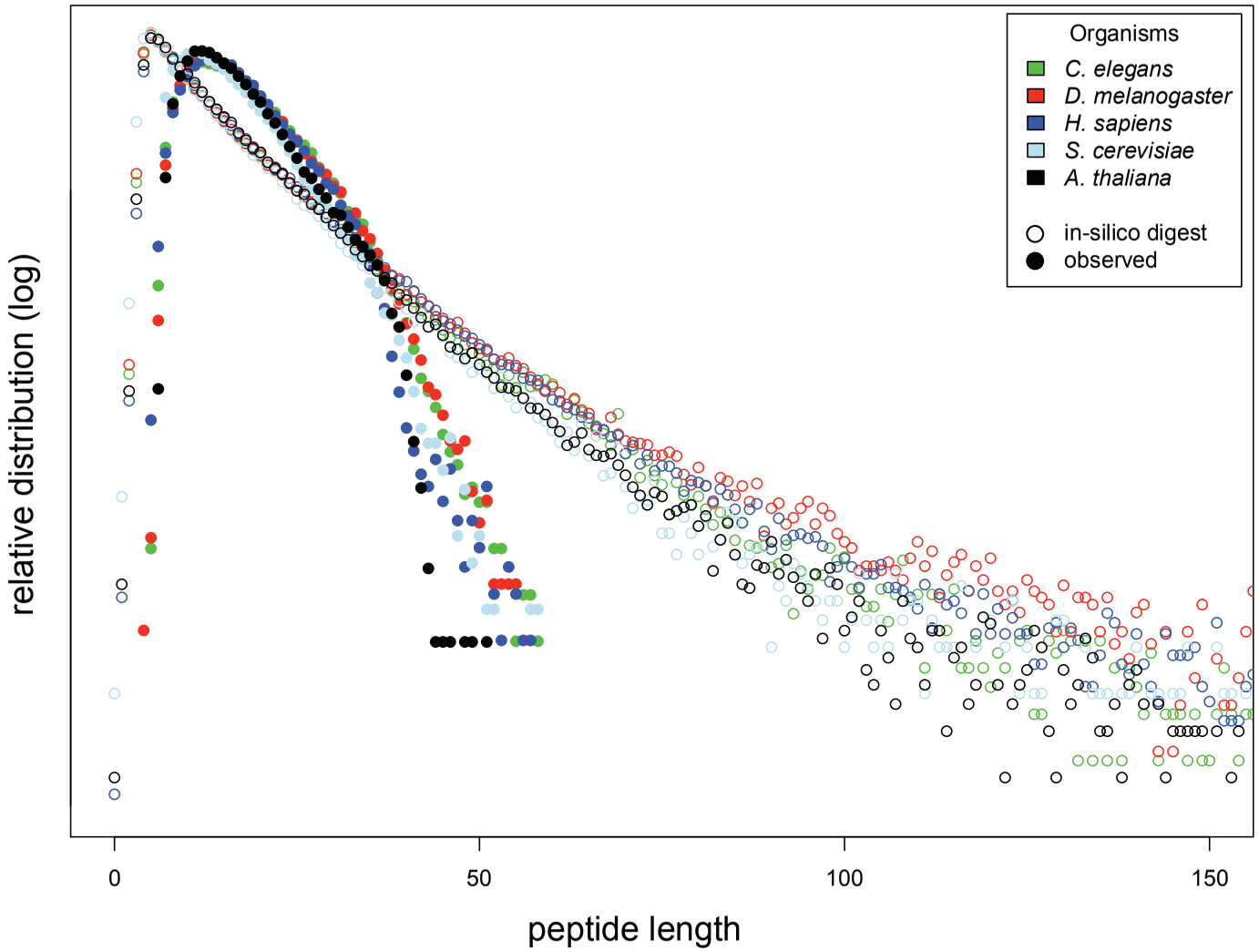
A. thaliana:
ThermoFinnigan LTQ

H. sapiens:
ThermoFinnigan LCQ Classic
ThermoFinnigan LCQ Deca
ThermoFinnigan LCQ Deca XP

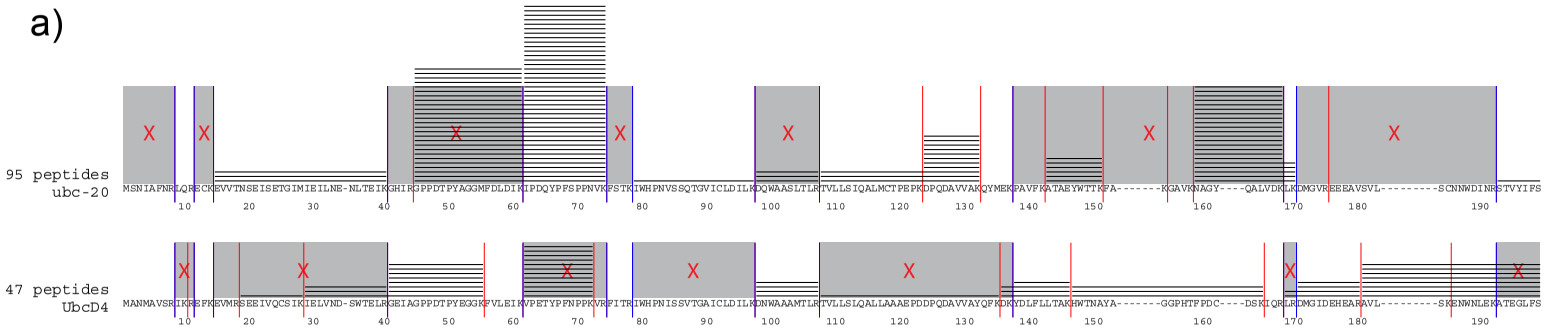
C. elegans:
ThermoFinnigan LTQ

D. melanogaster:
ThermoElectron LTQ

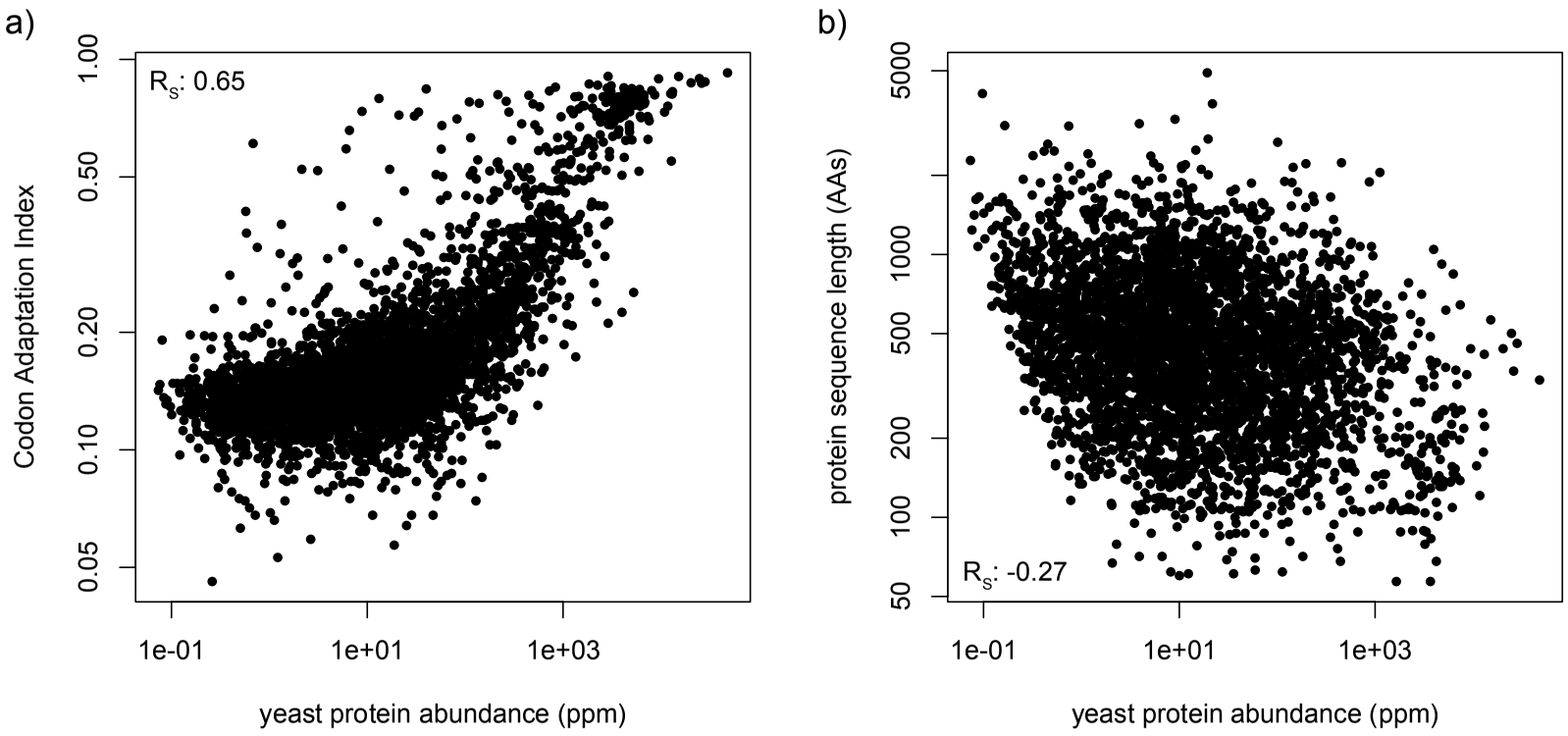
Suppl. Figure 1



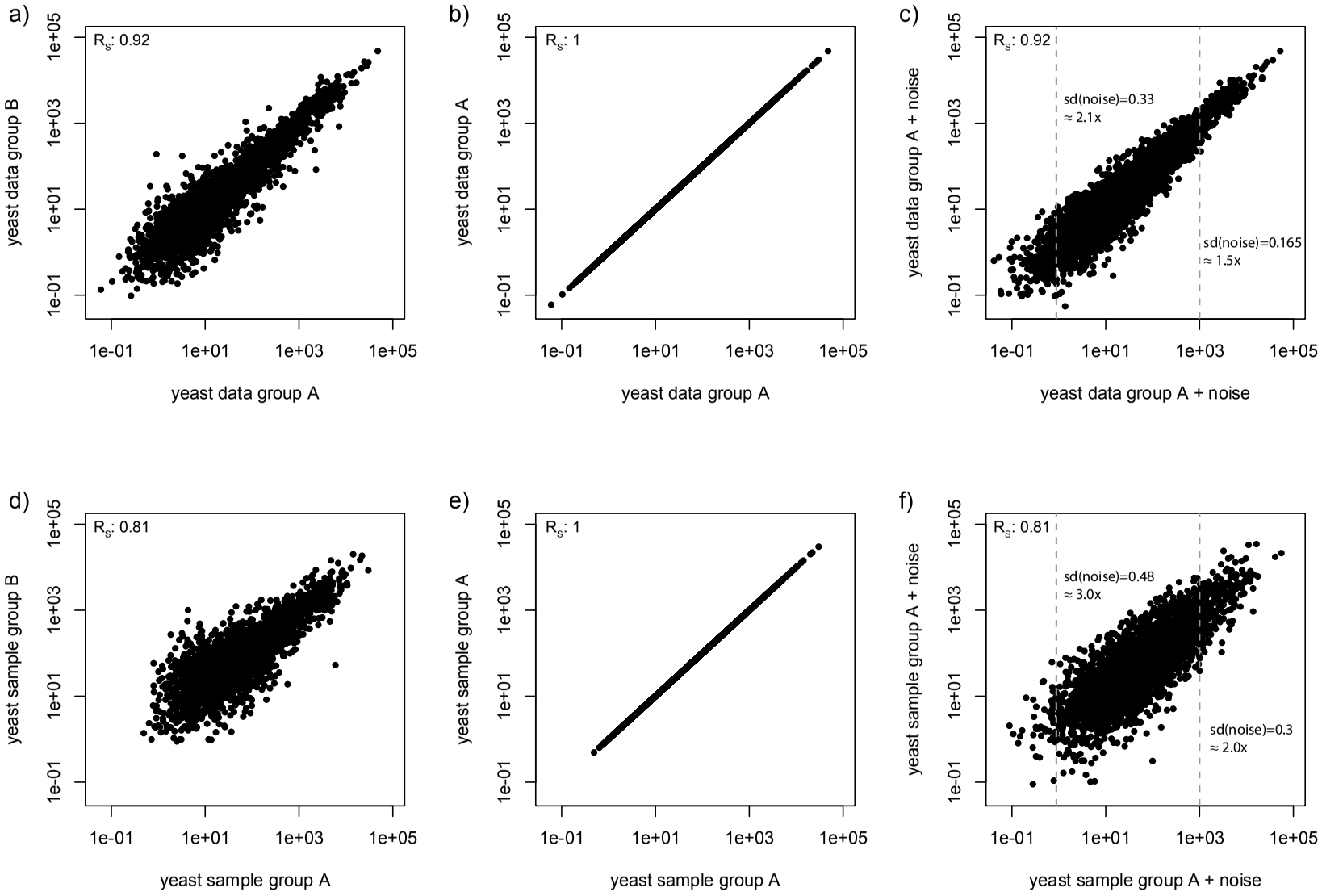
Suppl. Figure 2



Suppl. Figure 3



Suppl. Figure 4

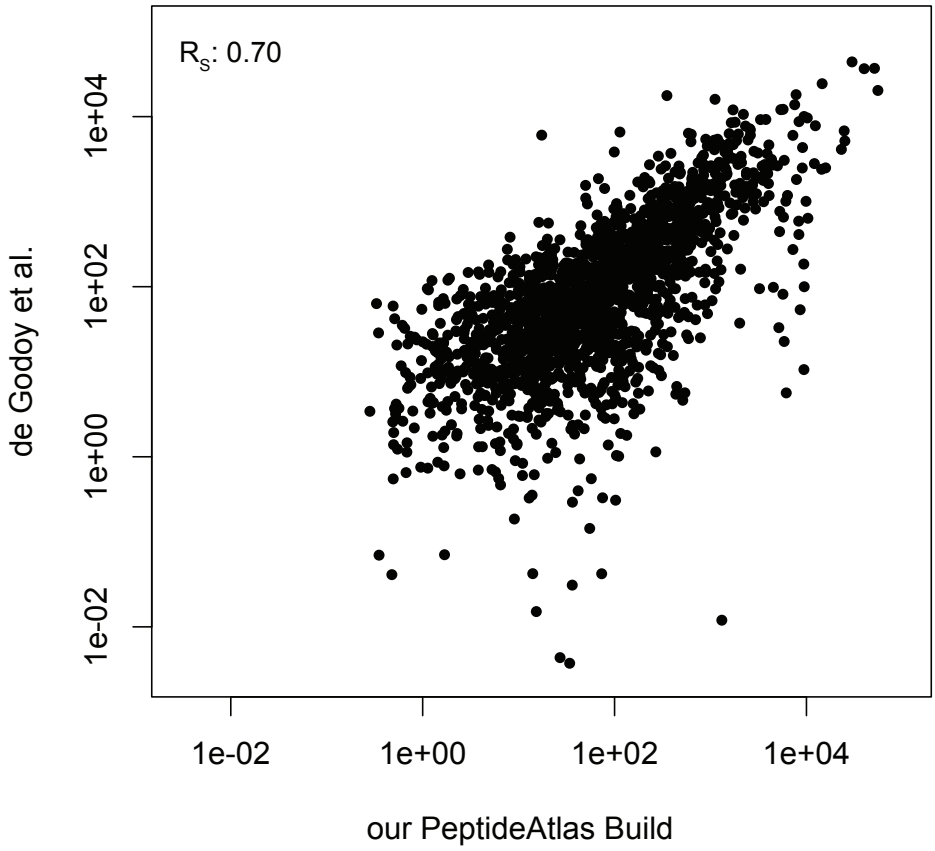


Suppl. Figure 5

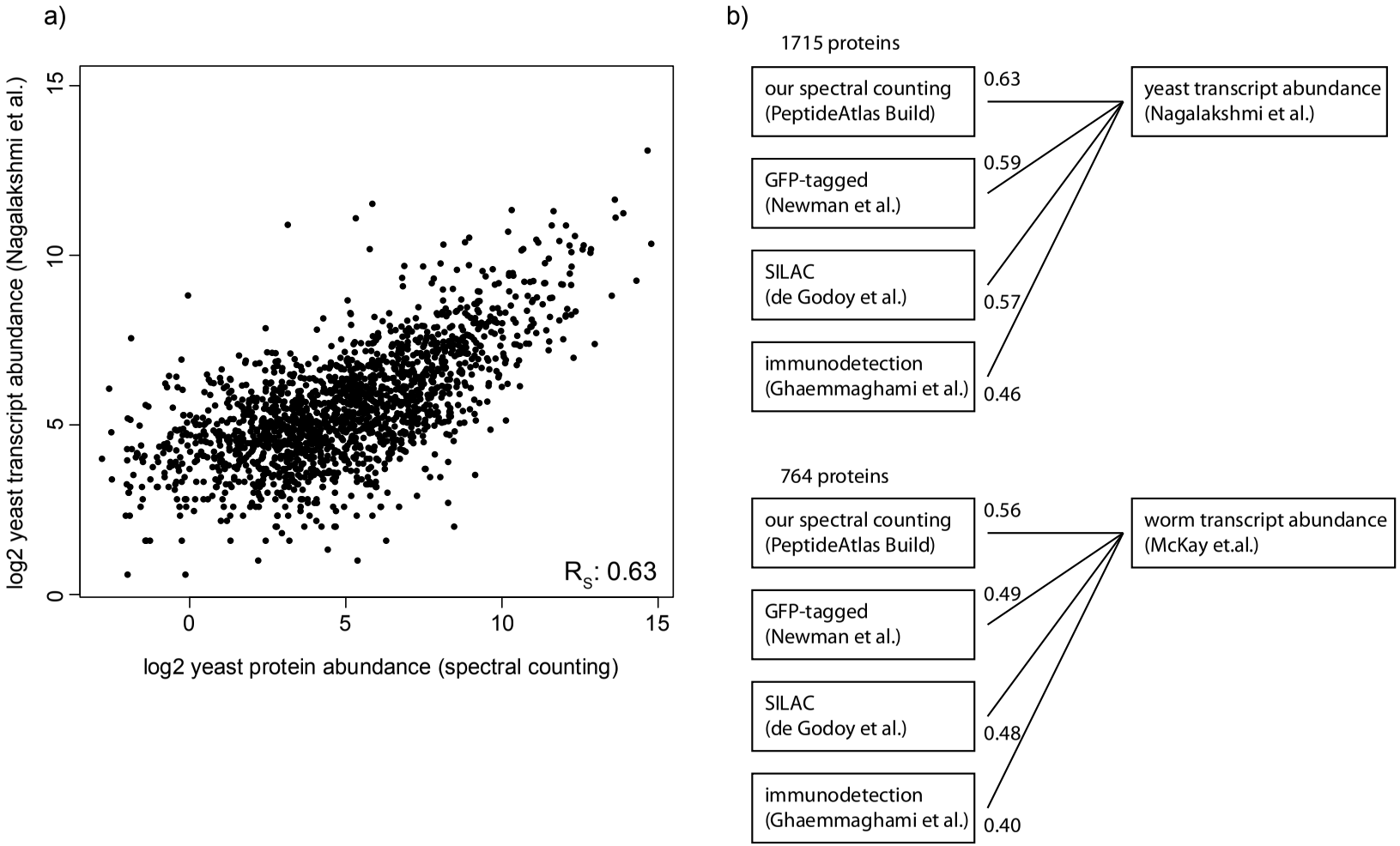
a)

Data set	Spearman's rho
our PeptideAtlas Build vs. Newman et al.	0.65
our PeptideAtlas Build vs. de Godoy et al.	0.70
our PeptideAtlas Build vs. Ghaemmaghami et al.	0.58
Newman et al. vs. de Godoy et al.	0.62
Newman et al. vs. Ghaemmaghami et al.	0.60
de Godoy et al. vs. Ghaemmaghami et al.	0.52

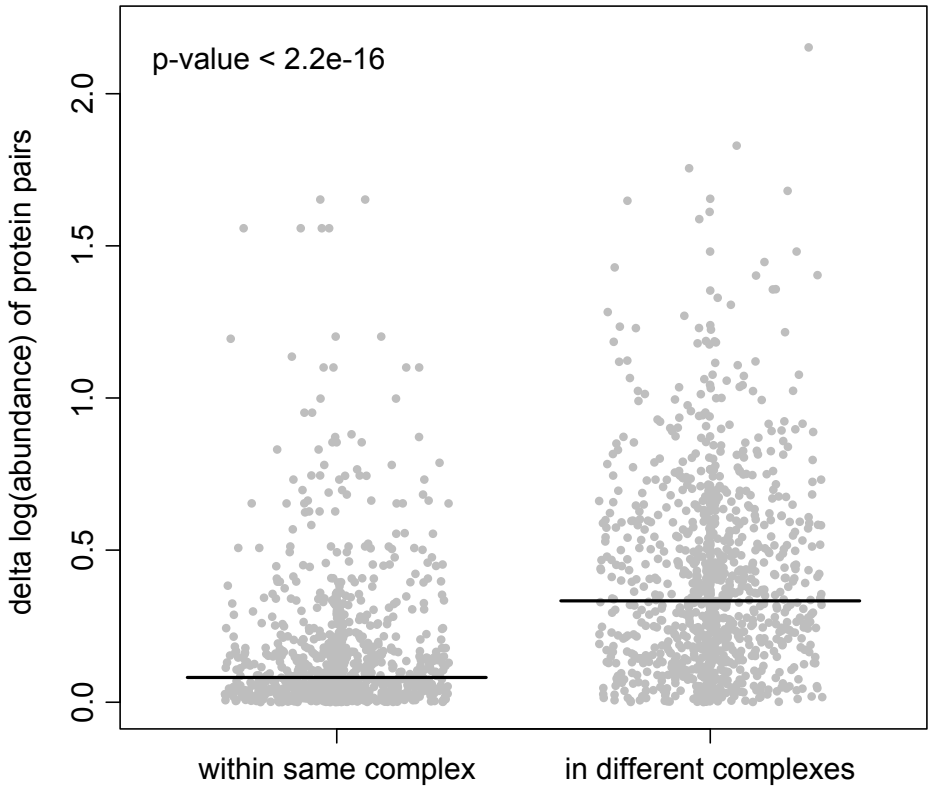
b)



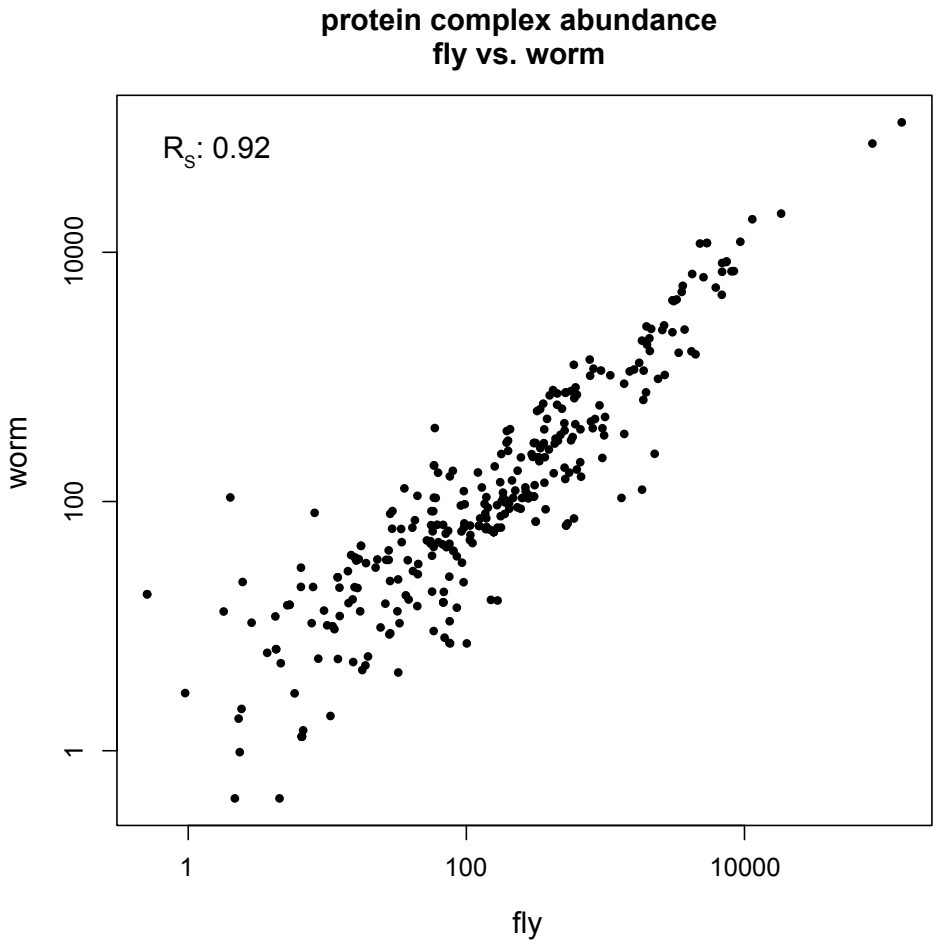
Suppl. Figure 6



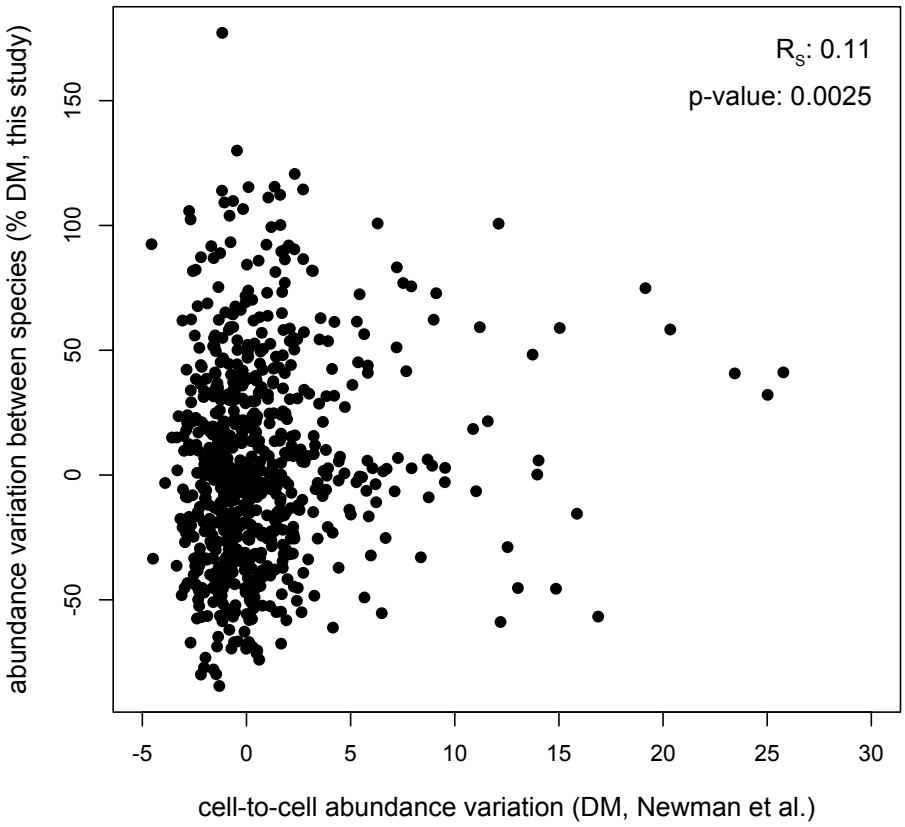
Suppl. Figure 7



Suppl. Figure 8



Suppl. Figure 9



Suppl. Figure 10

A) individual tests:

variable	given as	Spearman's rho	p-value
protein abundance	median of five organisms	-0.021	0.47
sequence conservation I	percent identity yeast ⇔ human	-0.061	0.037
sequence conservation II	percent identity fly ⇔ worm	0.034	0.30
gene duplications in yeast	number of duplicates in the orthologous group (yeast only)	0.060	0.04
gene duplications in five organisms	number of duplicates in the ortho-logous group (all five organisms)	0.160	3.2e-08
protein length	taken from yeast; in amino acids	-0.023	0.44
essentiality	as published for yeast; 0 or 1	-0.140	2.8e-06
cell-to-cell noise, poor medium	Newman et al.	0.110	0.0042
cell-to-cell noise, rich medium	Newman et al.	0.110	0.0031

B) independent, significant signals remaining in a generalized linear model:

variable	given as	Spearman's rho	p-value
gene duplications in five organisms	number of duplicates in the ortho-logous group (all five organisms)	0.133	0.0012
essentiality	as published for yeast; 0 or 1	-0.123	0.0086
cell-to-cell noise, poor medium	Newman et al.	0.178	0.0080

The dependent variable, for both tables, is **dm_percent** (i.e. protein abundance *variance* across organisms; the dependency of that variance with abundance itself has already been removed; this is the same variable as used in Fig 3 & 4 of the main text).

Suppl. Figure 11

6 Outlook

The instruments and methods in mass spectrometry-based proteomics are evolving fast, enabling researchers to measure proteins faster, with higher precision and reliability. From a rather exotic machine for biologists, the mass spectrometer is evolving into a standard tool that will soon be used by many wet lab scientists on a daily basis. Still, the correct analysis and interpretation of the output is not trivial. Bioinformaticians are struggling to keep up with the fast development cycle of the hardware and to deliver robust, easy to use analysis software that non-specialists can use.

We have refined and applied a simple approach, spectral counting, to infer quantitative information as a by-product from shotgun proteomics measurements. This procedure has been implemented as a software pipeline which computes abundance values based on a list of peptide counts and can be easily used by non-bioinformaticians.

For the first time, comprehensive abundance information acquired using a consistent method is available for several eukaryotic species. We could demonstrate that this information can be used to obtain interesting insights into the functional constraints on protein abundance evolution. Another interesting, and rather surprising, observation was the discovery that transcript abundances seem to be far less conserved across species than protein abundances. This implies a certain degree of freedom for transcript evolution, as changes in transcript abundance seem to be largely compensated by post-translational changes, e.g. translational efficiency or protein half-life, keeping the protein level stable. This has many implications for the evolvability of the quantitative composition of proteomes. It would also explain why correlations between transcript and protein abundances within one organism are usually reported to be relatively low.

As all these observations are very encouraging, we want to continue this line of work and extend the analysis to more organisms. This will allow us to study the evolution of protein abundances in more detail, focusing on how proteome composition depends on parameters like cell size, genome size, multicellularity/complexity or lifestyle/environment of an organism. With enough data, it should also be possible to identify scaling-laws and general principles that shape proteome evolution.

As more detailed data for higher organisms becomes available, the abundances could in principle also be computed separately for different tissues. Further, deviations from the 'normal' abundance of particular proteins in diseased tissues can serve to identify markers for diseases. In the very long run, we hope to identify families of proteins which are particularly sensitive to perturbations in expression levels and might therefore represent therapeutic targets.

Our first paper created quite some interest and several groups approached us to request our data for their own research. The group of Eugene Koonin at NCBI used our protein abundances for worm and fly to study the relative contributions of expression level and functional constraints to the rate of evolution of proteins. Their findings have recently been published in Genome Biology and Evolution (Wolf et al., Genome Biology and Evolution, 2010).

Our data was also used in another recent publication, exploring the relationship of expression levels and amino acid frequencies (Cherry et al., Molecular Biology and Evolution, 2010).

To continue with the work described in the previous two chapters, we want to apply our computational pipeline to more and more organisms as their proteome data become available. We want to make this protein abundance information available to the public and are therefore developing an online database, PaxDB (pax-db.org). This database will have a web interface for visualization of the data and will allow the user to browse, filter,

sort and download it in a variety of formats. It will also provide related information like orthologs, protein structure, functional annotations and protein-protein interactions imported from other databases.

7 Appendix

7.1 A Quantitative Targeted Proteomics Approach to Validate Predicted microRNA Targets in *C. elegans*

7.1.1 Preface

At the time of writing, this manuscript had been submitted to Nature Methods. I helped with the computational analysis and the proteome dataset.

A Quantitative Targeted Proteomics Approach to Validate Predicted microRNA Targets in *C. elegans*

Marko Jovanovic[1,2,3,4], Lukas Reiter[1,2,3,4], Paola Picotti[5], Vinzenz Lange[5,6], Erica Bogan[2,3], Benjamin A. Hurschler[7], Cherie Blenkiron[8,9], Nicolas J. Lehrbach[8,9], Xavier C. Ding[7], Manuel Weiss[2,3,4], Sabine P. Schrimpf[2,4], Eric A. Miska[8,9], Helge Grosshans[7], Ruedi Aebersold[5,6,10,*], Michael O. Hengartner[2,4,*]

1 contributed equally

2 Institute of Molecular Biology, University of Zurich, Zurich, Switzerland

3 PhD Program in Molecular Life Sciences Zurich, Zurich, Switzerland

4 Quantitative Model Organism Proteomics (Q-MOP), University of Zurich, Zurich, Switzerland

5 Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland

6 Competence Center for Systems Physiology and Metabolic Diseases, Zurich, Switzerland

7 Friedrich Miescher Institute for Biomedical Research (FMI), Basel, Switzerland

8 Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, United Kingdom

9 Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, CB2 1QN, United Kingdom

10 Faculty of Science, University of Zurich, Zurich, Switzerland

* Corresponding Authors:

Michael O. Hengartner
michael.hengartner@molbio.uzh.ch

Ruedi Aebersold
aebersold@imsb.biol.ethz.ch

Keywords: targeted proteomics, microRNA targets, *let-7*, *C. elegans*, selected / multiple reaction monitoring, SRM / MRM

Short Title: Quantitative Targeted Proteomics Approach to Validate microRNA Targets

ABSTRACT

Computational prediction methods for the identification of microRNA (miRNA) target genes face considerable challenges; in fact, the overlap in potential miRNA targets predicted for the nematode *Caenorhabditis elegans* by three commonly used algorithms is below 20%. Here we present a large-scale targeted proteomics approach to validate predicted miRNA targets in *C. elegans*. Using selected reaction monitoring (SRM), we quantified more than 160 proteins of interest in extracts from wild type and *let-7* mutant animals. We demonstrate by independent experimental downstream analyses such as genetic interaction, polysomal profiling and luciferase assays, that validation by targeted proteomics significantly enriches for biologically relevant *let-7* interactors. For example, we show that the zinc finger protein ZTF-7 is a *bona fide let-7* miRNA target. We propose that targeted mass spectrometry can be applied generally to validate candidate lists generated by computational methods or by large-scale experiments, and that the described strategy can easily be adapted to other organisms.

INTRODUCTION

miRNAs are short non-coding RNAs that bind to target mRNAs and negatively regulate gene expression. miRNAs play important roles in many developmental and disease-related processes[1,2]. A full understanding of miRNA function requires knowledge of their target mRNAs. One of the most widely used approaches to identify potential miRNA targets is to apply different target prediction algorithms (for review see Ref.1). However, the many algorithms available predict target sets with only limited overlap, e.g. below 20% for three available algorithms in *C. elegans*[3,4,5], and cumulatively identify several hundred potential target mRNAs per miRNA. In addition, lists generated from large-scale experiments undertaken to identify target mRNAs, such as studies based on mRNA profiling[6,7], pulldown of target mRNAs[8,9,10,11,12,13,14] and to a certain extent on genetics[15,16,17], have not been verified or reproduced by an independent large-scale method. Having a fast and conclusive experimental validation method to screen through a large set of potential miRNA targets and enrich for those with biological relevance would thus be of great value.

We reasoned that such a method should measure the most relevant output of gene expression, namely miRNA dependent changes in protein levels of the potential target genes. Moreover, in order to be worthwhile, the method should be easy to use, fast, sensitive, reproducible, quantitative and large-scale, as several hundred proteins have to be tested for each miRNA. A technique that promises to fulfill most of those criteria is proteomics. Indeed several groups have shown that shotgun proteomics can be successfully applied to screen for miRNA targets[18,19,20,21,22,23]. However, with available shotgun proteomics approaches, the bulk of measurement time is spent on signals not arising from the desired candidate proteins. Moreover, many of the desired proteins might not be measured due to the stochastic sampling of the peptide ions that is common to this method. This results in loss of sensitivity and reproducibility to a degree that high confidence data on candidate targets can only be achieved at a high cost of time and labor. In contrast, a targeted proteomics approach such as selected reaction monitoring (SRM)[24,25,26,27,28] has the potential for fast and reliable protein quantification of candidate

genes: By limiting the measurement to the proteins of interest, the sensitivity and the reproducibility of the measurements increase dramatically. This can be achieved by selecting for each candidate miRNA target protein one or several proteotypic peptides (PTPs) - peptides that unambiguously identify a protein of interest and have favorable detection properties by mass spectrometry[29,30,31]. A wide range of quantification methods for proteomics are available[32,33,34,35,36]. One of the first widely used quantification method is based on isotope-coded affinity tags (ICAT). In addition to being a robust relative quantification method, ICAT also reduces sample complexity due to labeling and enrichment of cysteine containing peptides only[33].

We recently published a large *C. elegans* proteomics dataset (*C. elegans* proteome atlas), in which 8608 proteins, or about 40% of the proteome was identified by shotgun proteomics experiments[37,38]. This large dataset provides the basis to retrieve the necessary PTPs to conduct a targeted proteomics experiment in *C. elegans*.

Here, we describe the application of SRM-based targeted proteomics and ICAT quantification to screen scores of potential *let-7* targets in *C. elegans*. Our targeted proteomics approach provided high confidence quantification data, which we then mined to identify potential miRNA targets of biological significance. Independent downstream experiments, including genetic studies, polysomal profiling and luciferase assays, confirmed that the candidate genes classified as regulated by *let-7* based on our protein quantification data are indeed enriched in *let-7* interactors. Based on this proof of principle study, we suggest that SRM-based targeted proteomics can be widely used to screen candidate lists generated by computational methods and/or by large-scale experiments.

RESULTS

Exact protein quantification in *C. elegans* using targeted proteomics

To test if a targeted proteomics approach is suitable for *C. elegans* whole animal extracts, we tested SRM for mass spectrometry (MS) measurements in combination with ICAT sample labeling for relative quantification. We prepared a 1:1 mixture of heavy and light ICAT-labeled extracts from a mixed stage population of wild-type *C. elegans* animals, selected 5 proteins of different abundance classes based on our *C. elegans* proteome atlas[37] and measured their abundance ratio. We measured at least two PTPs per protein and the mean value for the light:heavy ratios of all the measured peptides was 0.97 (expected ratio 1:1), with a relative standard deviation of 15.5% (Supplemental Figure 1). Moreover, the ratio of different PTPs for the same protein were all in good agreement, independent of absolute signal intensity. This experiment showed that our setup allows for exact quantifications of proteins of interest in a complex whole animal extract generated from *C. elegans*.

SRM-based validation of potential *let-7* target genes

As a proof of principle, we applied the SRM/ICAT proteomics method to screen through a set of several hundred potential *let-7* miRNA targets. We focused on *let-7* because it is highly conserved from *C. elegans* to humans[39] and is one of the best studied nematode miRNAs[16,17,40,41,42,43,44,4]. We used for our studies the hypomorphic allele *let-7(n2853)*, which contains a point mutation within the mature *let*-7 seed sequence that also results in a reduction in *let*-7 expression levels [16].

The experimental strategy used to quantify potential *let-7* targets is described in Figure 1. Briefly, we first compiled a list of potential *let-7* targets based on predictions from five different algorithms[4,3,5,17,45], experimental data (e.g. microarray analysis, RNAi screens, etc.) and the published literature, including known *let-7* target genes[16,17,40,41,42,43,44,46,47]. We also included control genes that we knew to be altered in *let-7* mutant animals due to secondary effects (Hurschler B *et al.*, unpublished data) and randomly chosen genes which served as "neutral controls", resulting in a final list of 861 candidate genes

(Supplemental Table 1) of which 650 were present in the *C. elegans* proteome atlas. For 391 of these, we had observed cysteine-containing peptides, a prerequisite for applying ICAT quantification. We could experimentally confirm the presence of 181 (46%) of these 391 proteins by SRM-triggered MS/MS measurements in extracts from synchronized L4 larvae. These results established the basis for the reliable quantification of 181 potential target proteins in *C. elegans* extracts.

We next compared the abundance of these 181 proteins in synchronized *let-7(n2853)* mutants and wild-type late L4 larvae (when *let-7* miRNA expression is the highest) in three biological replicates. Most target proteins (139) could be quantified in all three biological replicates, another 15 in two replicates and 7 in one replicate, yielding quantification data for a majority of the identified proteins (161 out of 181; 89%), confirming the high reproducibility of this method (Figure 2a and Supplemental Table 2). The normalized log2 ratios (*let-7(n2853)* versus wild-type) and corresponding *P*-values for all 161 proteins that we quantified are shown in Figure 2b and listed in Supplemental Table 2.

Twenty nine proteins showed a significant difference in expression level between *let-7(n2853)* and wild-type animals (*P*-value < 0.01, one sample Student's t-test; Figure 2c and Supplemental Table 3). Of these, 10 were down- and 19 were upregulated in *let-7(n2853)* animals. As expected, the two control genes *vit-2* and *vit-6*, which show a greatly reduced abundance at the mRNA level in *let-7(n2853)* mutants were also strongly downregulated in our assay (13-fold and 23-fold respectively; Supplemental Table 3). The remaining 27 regulated proteins included LET-526 (also known as LSS-4), the only previously reported *let-7* target whose abundance we could measure (see below), 15 of the 66 predicted *let-7* targets based on bioinformatics and 9 of 53 proteins, whose mRNA do not contain a predicted *let-7* target sites, but that have been linked to *let-7* through other experimental approaches or the literature (Supplemental Table 3). By contrast, only 2 out of 39 of the randomly picked "neutral controls" showed a significant abundance change. Interestingly the "neutral controls" were the only significantly underrepresented group among the regulated proteins (*P*-value = 0.016, Fisher's Exact Test). This low "hit

rate" for these randomly tested proteins confirms that our initial candidate list was indeed enriched for *let-7* miRNA target genes.

Whether the regulated candidates are primary or secondary targets of *let-7* cannot be determined from the protein ratios. Although the most straightforward explanation for proteins downregulated in *let-7(n2853)* mutants is secondary effects, as it is the case for the two control proteins VIT-2 and VIT-6 (Hurschler B *et al.*, unpublished data), miRNAs have recently been reported to act as positive regulators under certain conditions[48]. A gain-of-function caused by the point mutation within the seed region of the mature *let-7* miRNA in *let-7(n2853)*, resulting in better binding to a suboptimal seed sequence, also cannot be excluded at this point.

LET-526 shows a splice-variant specific response to *let-7*

Previous studies had identified eight *let-7* miRNA targets[46,47]. We could quantify one of these in our assay, namely C01G8.9, also known as *lss-4* or *let-526*[Ref.17]. As expected, we observed significant upregulation of LET-526 protein levels in *let-7(n2853)* mutants (Supplemental Table 3). However, we noticed that the two measured peptides matching to this protein showed a strong discrepancy in the strength of regulation. Interestingly, this discrepancy correlated with the known alternative splicing pattern of LET-526. Whereas the peptide specific for the LET-526a splice form showed a strong, 3.1-fold change, the peptide matching to a region common to both splice-isoforms displayed only a weak 1.2-fold upregulation in *let-7(n2853)* mutant animals (Figure 3a, b).

To verify this splice-variant specific response through an independent experimental approach, we determined the extent of polyribosome association of the LET-526a and LET-526b mRNAs in L4 staged wild-type and *ain-2(RNAi); ain-1(ku322)* double mutant animals. The GW182 proteins AIN-1 and AIN-2 are required for miRNA function[49,50,8] and known miRNA targets display a shift towards the highly translated polysomal fractions in *ain-2(RNAi); ain-1(ku322)* mutants relative to wild-type animals due to the lack of miRNA mediated translational repression[49].

We found a strong shift of the *let-526a* mRNA towards the polyribosome fractions upon AIN-1/AIN-2 depletion (*P*-value = 0.03, one-sided Student's t-test). By contrast, probes

detecting both splice variants failed to detect a statistically significant shift (*P*-value = 0.17, one-sided Student's t-test; Figure 3c). Taken together, our results suggest that the *let-526a* mRNA responds much more strongly to *let-7* activity than the *let-526b* isoform. The different response of the two splice-variants to *let-7* misexpression is intriguing because based on EST data, both splice variants have the same 3'UTR and would therefore be expected to contain the same predicted *let-7* binding sites[17]. Whether the b isoform is resistant to *let-7*-mediated repression or whether it is expressed in a different set of tissues than *let-7* remains to be determined.

Validation by targeted proteomics significantly enriches for *let-7* genetic interactors

The aim of this work was to develop a proteomics-based validation method to select from a long list of genes the biologically relevant candidates that warrant a more detailed downstream analysis. *let-7(n2853)* mutant animals grown at 25°C die at the adult stage due to vulval bursting. Knockdown by RNAi of known *let-7* miRNA targets has been shown in some cases to rescue this lethality to different extents[17,43,42,44,4]. To determine whether the positive hits in our proteomics screen are indeed enriched in *let-7* targets, we knocked-down all 29 genes that showed protein changes in *let-7(n2853)* mutants (up- or downregulated) to determine whether they could suppress the *let-7* lethal phenotype. Six of the 29 genes knocked down by RNAi caused either larval arrest or lethality, and thus could not be scored for suppression of vulval bursting. From the remaining 23 genes tested, ten successfully and reproducibly rescued the lethality to at least 20% (Figure 4a and Supplemental Table 4). As a control, we performed a similar experiment using 29 candidate genes that did not show significant protein changes in our targeted proteomics assay. Again, five genes either caused early larval arrest or lethality and could not be characterized further. Only three out of the remaining 24 candidates were able to rescue the lethality (Figure 4b and Supplemental Table 5), demonstrating that the regulated protein set is significantly enriched for genes that genetically interact with *let-7* (*P*-value = 0.024, Fischer's Exact Test; Figure 4c). We conclude that a targeted proteomics method can indeed be used to enrich for miRNA interaction partners.

Protein abundance changes are only partially recapitulated at the mRNA level

In addition to causing translational repression, miRNAs can also lead to degradation of their targets[1]. To determine whether the changes in protein levels that we observed could also be captured at the mRNA level, we determined the mRNA levels for all 161 proteins by RT-qPCR in *let-7(n2853)* and wild-type animals (Supplemental Table 6). Plotting protein changes as a function of mRNA changes revealed that while some proteins showed a very good correlation between mRNA and protein changes, others, including the known *let-7* target *let-526*, showed a significant protein change but no strong change in mRNA levels (Figure 5a). We also specifically looked at the 47 genes that were scored for suppression of *let-7* lethality (regulated candidates and non-regulated controls - see above). Interestingly, whereas many of the 13 RNAi suppressors showed large changes in protein levels in *let-7(n2853)* mutants, their mRNA levels varied only weakly if at all (Figure 5b). We conclude that many of the protein changes we detected in our targeted proteomics approach are not recapitulated on the mRNA level, and that while mRNA profiling can yield significant results, it would fail to detect several of the biologically important candidates provided by protein quantification.

The zinc finger protein ZTF-7 is a new *bona fide let-7* miRNA target

One of the most interesting candidates from our RNAi screen was *ztf-7* (F46B6.7), as knockdown of this gene not only suppressed lethality (see Figure 4 and Supplemental Table 4), but also the sterility observed in *let-7(n2853)* mutants at 25 °C (data not shown). Of the genes that we tested, only the two positive controls *daf-12* and *lin-41* – both well established *let-7* targets – could also suppress both defects. In order to confirm the suppression of lethality by RNAi against *ztf-7* in *let-7(n2853)* animals, we crossed *ztf-7(tm600)* mutant animals with *let-7(n2853)* mutants and tested the double mutant for suppression of lethality at 25°C. Indeed lethality was strongly reduced in *ztf-7(tm600); let-7(n2853)* double mutant animals when compared to the *let-7(n2853)* single mutants (Figure 6).

We had quantified ZTF-7 in all three biological replicates and based on our targeted proteomics measurements, ZTF-7 protein levels were elevated in *let-7(n2853)* mutants when compared to wild-type animals. Although overall ZTF-7 protein levels were up by

only 10% in *let-7(n2853)* worms, this upregulation was highly reproducible and significant (*P*-value = 0.005, one sample Student's t-test; Figure 6).

As *ztf-7* is predicted to contain at least one conserved *let-7* binding site[4,51,5], we next tested whether the *ztf-7* 3'UTR is able to confer *let-7*- dependent regulation of a reporter transcript. It has been reported that certain *C. elegans* 3'UTRs can elicit a miRNA dependent response in human cell lines[52]. As the sequence of the mature *let-7* miRNA is identical in *C. elegans* and in humans[39], we could rapidly test the effect of both overexpression and depletion of human let-7a miRNA in HeLa cells, which were transfected with a dual luciferase vector where the *ztf-7* 3'UTR was cloned directly downstream of the firefly luciferase gene (luciferase::ztf-7 $^{3'\ UTR}$). Indeed, we observed a strong response of the luciferase::ztf-7 $^{3'\ UTR}$ reporter to both human let-7a up- and downregulation (Figure 6).

Taken together, our proteomics, genetic and reporter assays strongly suggest that *ztf-7* is a *bona fide let-7* miRNA target. Further work will be required to understand the function of ZTF-7 in *C. elegans* development.

DISCUSSION

Many computational and experimental large-scale approaches have been developed to identify miRNA target genes[1,18,19,13]. However, most of those approaches yield hundreds of potential targets, and it is difficult to separate the false positives from the true positive targets. Therefore, an independent high throughput validation method would be extremely beneficial. We describe here the development of such an experimental validation pipeline, which is based on targeted proteomics.

In a first step, we showed that SRM in combination with the ICAT quantification method yields reproducible, exact and precise relative protein quantification results for selected proteins in a complex *C. elegans* extract. We then applied our targeted proteomics approach to screen a large list of potential *let-7* targets for significant protein changes upon *let-7* perturbation. Approximately half of all our candidates could be quantified. After statistical analysis of the data, 29 out of 161 quantified candidates showed significant expression changes in *let-7(n2853)* mutant animals. All the positive controls and the known *let-7* target quantified were among those 29 candidate genes, proving the suitability of our approach. Additionally we could demonstrate that the 29 candidates are significantly enriched in genetic interactors with *let-7*. Several of the biologically relevant *let-7* interactors showed no strong change at the mRNA level, and thus would have been missed by mRNA profiling alone. Finally, we established the zinc finger protein ZTF-7,as a *bona fide let-7* target, showing successfully that our validation tool can lead directly to the identification of target genes. *ztf-7* was also identified recently as a potential *let-7* target by Andachi, using a novel method to identify miRNA target genes[53], thus providing independent support for our claim that *ztf-7* is a *bona fide let-7* target.

Our results demonstrate the suitability and advantages of a targeted proteomics approach to find biologically relevant candidate miRNA targets. First, this method measures changes in protein levels, arguably the most relevant assay for miRNA activity. Second, our approach allows for the quantification of several hundred proteins and thus has a much higher throughput than traditional protein quantification methods such as immunoblotting. Additionally the development of suitable mass spectrometric assays is

significantly faster and cheaper than of immuno assays[54]. Moreover, once an SRM assay is established for a protein, it becomes universally useful and exportable[55,28]. We already established such a public database of validated SRM assays for approximately 1500 Saccharomyces cerevisiae proteins[28]. By the same token, our established SRM assays for the 181 *C. elegans* proteins measured are also of universal use (Supplemental Table 7). Third, because it focuses on highly responsive peptides, our SRM-based approach is highly sensitive and reproducible. Indeed, we could reproducibly measure changes as small as ten percent in total protein abundance, as exemplified with ZTF-7. This high accuracy is particularly important in the analysis of potential miRNA targets, as miRNAs have been suggested to mostly induce small changes in target gene expression[18,19].

By choosing the *let-7* miRNA as our "test candidate", we in fact challenged the sensitivity of our approach even further, as the *let-7* miRNA is not expressed in the whole animal[56]. Thus, changes in protein levels of targets that are co-expressed with *let-7* in only a few cells of the animal might be masked by the stable expression of the protein in the rest of the animal, where *let-7* is not present. Indeed it is for example very likely that *ztf-7* is regulated by *let-7* only in a subset of tissues where it is expressed. Since biologically significant candidates were identified even in such a challenging situation, the application of the method to more homogenous conditions, such as human cell lines, should be straightforward.

Despite the clear value of our targeted proteomics approach, several challenges remain. First, the processing of the raw SRM data is still cumbersome – quantifications are not yet automated and the assignments of the correct peak groups to their corresponding peptides is not always straightforward and therefore has a low percentage error associated to it. Both issues can likely be solved in the future through further algorithm development[57]. Second, the targeted proteomics method described here is based on the ICAT quantification strategy, which limits quantification to cysteine-containing peptides. Unfortunately, the majority of PTPs contain no cysteine[37]. However, our approach can readily be adapted to other quantification strategies[27]. For example, the use of heavy isotope-labeled worms would allow access to the full repertoire of *C. elegans* PTPs[34].

Third, although we achieved a high sensitivity, we could quantify only approximately 50% of all the proteins we had on our final target list. Many of the proteins that we could not measure are probably not expressed at the late L4 stage. For other proteins, our current protocol was possibly not sensitive enough. A strategy to increase sensitivity is to use chemically synthesized peptides to optimize assays. This pre-optimization step has been shown to enable quantification of even the lowest abundance proteins in yeast[27]. Finally, a biological limitation of this targeted proteomics approach is that we are unable to distinguish primary from secondary targets. Additional experiments will invariably be necessary to establish which hits are direct targets, as has been exemplified for *ztf-7* (Figure 6).

The targeted proteomics approaches described here should be considered complementary to the shotgun proteomics approaches recently reported to identify miRNA targets[58,18,22,21,20,23,19,59]. While shotgun proteomics should be regarded as one of several discovery tools that can be used to find potential new miRNA target candidates, a targeted proteomics approach should be perceived as a validation / hypothesis driven tool with high sensitivity, reproducibility and accuracy.

Importantly, while we applied the targeted proteomics method described here for the validation of miRNA targets in *C. elegans*, the method is broadly applicable, and can readily be adapted to other organisms and to other biological questions. A wide range of quantification methods are available[32,33,34,35,36], suitable for nearly every extract composition. In addition, there are public proteomics databases for a wide range of different organisms, including *Drosophila melanogaster*[60,61], humans[62] and *Arabidopsis thaliana*[63], where experimentally identified proteins and their corresponding PTPs can be easily mined. Even for organisms where such proteomics data is not readily accessible, sophisticated PTP prediction algorithms[30,31] can be consulted in order to target the right peptides. Thus, the targeted proteomics approach described here can be applied generally to measure protein abundance of long candidate lists generated by computational methods or by large-scale experiments.

ACKNOWLEDGEMENTS

We thank Alexander Stark for sharing miRNA target predictions for *C. elegans*. We also thank Martin Moser for the assistance with the RT-qPCR assays; Bernd Roschitzky and Bertran Gerrits for technical support; Hubert Rehrauer for statistical support; Ralph Schlapbach for access to the Functional Genomics Center Zurich; the Hengartner, Aebersold, Grosshans and Miska laboratories, Erich Brunner and the whole Q-MOP team for insightful discussion and comments on the manuscript.

This work was funded in part by the University of Zurich Research Priority Program in Systems Biology/Functional Genomics, the Swiss National Science Foundation, the GEBERT RÜF Foundation, SystemsX, the Ernst Hadorn Foundation and the ETH Zurich. M.J. and L.R. were supported by a grant from the Research Foundation of the University of Zurich. M.J. was also supported by a fellowship from the Roche Research Foundation. P.P. is the recipient of a Marie Curie Intra-European fellowship. V.L. was supported by the Competence Center for Systems Physiology and Metabolic Diseases. H.G. was supported by the Swiss National Research Foundation and the Novartis Research Foundation and X.C.D by a Boehringer Ingelheim Funds PhD Student fellowship. C.B., N.J.L and E.A.M. were supported by a Cancer Research UK Programme Grant to E.A.M.

AUTHOR CONTRIBUTIONS

M.J., L.R., M.O.H. and R.A. designed the experiments and wrote the paper. L.R. and M.J. did the majority of the data analysis. M.J. did the majority of the experiments. P.P. and V.L. contributed to and supervised the SRM experiments. E.B. contributed to the RNAi and RT-qPCR experiments. B.A.H. and X.C.D. performed the polysomal profiling experiments. C.B. and N.J.L. contributed to the reporter assays. S.P.S. and M.W. provided the *C. elegans* proteome atlas. H.G. and E.A.M. provided critical input on the manuscript, contributed significantly to the experimental design, the data and the data analysis. M.O.H. and R.A. supervised the whole project.

FIGURE LEGENDS

Figure 1: Strategy and workflow for targeted protein quantification.

(a) General proteomic strategy and (b) workflow for the quantification of potential *C. elegans let-7* interacting genes. (a) Proteins of interest were compiled based on literature and previous experiments. PTPs for these proteins of interest were selected from the *C. elegans* proteome atlas[37]. The selected PTPs were used as probes for reproducible quantification by SRM on a QTrap mass spectrometer operated as a triple quadrupole instrument. (b) From the initial set of 861 proteins, 650 had PTPs in the *C. elegans* proteome atlas, of which 391 had cysteine containing peptides and were quantifiable by ICAT. Validated transitions were derived for 181 proteins, of which 161 could be quantified. Of these, 29 proteins showed significant changes in abundance between wild type and *let-7(n2853)* mutants.

Figure 2: Identification of proteins regulated by the *let-7* miRNA.

(a) Heat map and hierarchical clustering of the 161 quantified proteins in three separate biological replicates (b.r.1 – b.r.3). Red and blue indicate up- and downregulated proteins in the *let-7(n2853)* mutant, respectively (see color code). (b) Volcano plot: normalized mean log2 ratios and probability of regulation (-lg (*P*-value)) of the measured proteins. Predicted *let-7* targets[4,3,5,17,45] are shown in orange, all other proteins in blue. All proteins above the dotted red line (at *P*-value = 0.01) were considered to be significantly regulated. (c) Heat map and hierarchical clustering of the 29 significantly regulated proteins (*P*-value < 0.01).

Figure 3: Splice variant-specific regulation of the *let-7* miRNA target let-526/lss-4

(a) Genomic structure of the *let-526* (C01G8.9) locus. The two gene models, *let-526a* and *let-526b* are depicted. Black boxes represent coding exonic sequences and grey boxes the untranslated regions (5'UTR and 3'UTR). EST evidence was used to map the 3'UTR to LET-526b as shown. The two peptides that were quantified are indicated in red. Peptide LIEFCEHNGEPLTMVPQVSK is unambiguous and specific for LET-526a only, while peptide VPEATDSSIPCPVSPR is ambiguous and cannot distinguish between LET-526a

and LET-526b. (b) Representative LC-SRM chromatograms showing the SRM measurements of the two peptides from *let-7(n2853)* and wild-type extracts. Each peptide was measured using two transitions. For transition 1, the red and grey lines correspond to the signal intensities in *let-7(n2853)* extracts and wild-type extracts, respectively. For transition 2, *let-7(n2853)* intensity is depicted in blue and wild-type in green. The fold change ratios (*let-7(n2853)* versus wild-type) averaged over all SRM measurements and over all biological replicates for each peptide and the corresponding standard deviations are shown within the respective chromatograms. (c) *let-526a* and *let-526a,b* mRNA distributions, determined by RT-qPCR, across polysomal profiles of L4 stage wild-type and *ain-2(RNAi); ain-1(ku322)* mutant worms. The *let-526a,b* RT-qPCR primers detect both splice variants, whereas the *let-526a* RT-qPCR primers are specific to the *let-526a* isoform. The dotted black line indicates the boundary between monosomes and polysomes. Representative polysome profiles of wild-type and *ain-2(RNAi); ain-1(ku322)* mutant worms are shown above. Polysomal profiling experiments were performed in triplicate. The error bars depict the standard error of the mean.

Figure 4: Genes displaying protein changes in *let-7(n2853)* mutant animals are enriched in *let-7* suppressors.

let-7(n2853) animals grown at 25°C die at the adult stage due to vulval bursting. Knock-down of some known targets has been shown to rescue this lethality to different degrees[17].

(a) The 29 genes that showed protein changes were knocked-down by RNAi to determine if they suppress the *let-7* lethal phenotype. Less than 5% of the *let-7(n2853)* animals treated with control RNAi (vector RNAi or ZK617.1(*unc-22*) RNAi) survived as adults. Six genes could not be scored, as their RNAi inactivation led to either lethality or larval arrest. The remaining 23 candidates are represented by the grey bars, and the positive controls (F11A1.3 (*daf-12*), C12C8.3 (*lin-41*) and C18D1.1 (*die-1*) RNAi) are depicted as the black bars. Only survival rates above 5% are shown. (b) As a control, 29 candidates that did not show a significant protein change in the *let-7(n2853)* mutant animals in our targeted proteomics assay were tested as in (a), including the same positive and negative controls. Again, 5/29 genes could not be scored due to lethality or larval arrest. (c) The

"regulated" candidates are significantly enriched in *let-7(n2853)* suppressors compared to the "not regulated" group. The *P*-value of enrichment (Fisher's Exact Test) was calculated for different survival cutoffs. The table lists the number of suppressors for each group ("regulated" and "not regulated") at the listed cutoffs and the corresponding *P*-values.

All the experiments were performed at 25°C and in triplicate. The error bars depict the standard error of the mean. The candidate marked by the asterisk (*) showed suppression in two out of three replicate experiments and was regarded as positive as the average survival rate over all three replicates was above the threshold.

Figure 5: Comparison of *let-7*-dependent changes in protein and transcript levels of candidate *let-7* miRNA targets.

log2 fold changes at the mRNA (x-axis) and protein (y-axis) level between *let-7(n2853)* mutant and wild type worms for (a) all 161 proteins measured and (b) all 47 candidates scored by RNAi. (a) Proteins that are predicted *let-7* targets[4,3,5,17,45] are shown in orange, all other proteins in blue. (b) Proteins that showed suppression of lethality are shown in red; proteins that did not show suppression of lethality are shown in blue. The error bars indicate the standard error of the mean over the three biological replicates.

Figure 6: ZTF-7 (F46B6.7) is a *bona fide let-7* target

(a) ZTF-7 (F46B6.7) protein is significantly upregulated in *let-7(n2853)* mutants compared to wild-type animals. The average fold change (*let-7(n2853)* / wt) over all measurements and the calculated *P*-value (Student's One Sample t-test) are shown. (b) *let-7(n2853)* animals grown at 25°C die at the adult stage due to vulval bursting. Knock-down of some known targets has been shown to rescue this lethality to different degrees[17]. 3% of *let-7(n2853)* mutants are still alive 12 hours post L4. In contrast 43% of *ztf-7(tm600); let-7(n2853)* double mutants are alive 12 hours post L4. *ztf-7(tm600)* mutants did not show any lethality 12 hours post L4 and no other obvious defects. All the experiments were performed at 25°C and in quadruplicate. The error bars depict the standard error of the mean. (c) Relative luciferase activities for reporter constructs containing the indicated 3'-UTR sequences. The let-7a readouts (mimics and inhibitors)

were normalized to their respective oligo controls (see Materials and Methods for details). The 3'UTRs of the known targets C12C8.3 (*lin-41*), F11A1.3 (*daf-12*) and F13D11.2 (*hbl-1*) were included as positive controls while the 3'UTRs of F36A4.7 (*ama-1*) and T04C12.6 (*act-1*) were added as negative controls[43,42,44,17]. Transfections were performed in triplicate for all candidates but *lin-41* (marked by asterisk (*)), which was only transfected in duplicate.

MATERIALS and METHODS

Mutations

All mutants used in this study were derived from the wild-type variety Bristol strain N2. The following mutations were used: LGV: *ztf-7(tm600)* (see http://www.wormbase.org); LGX: *let-7(n2853)*[16], *ain-1(ku322)*[50].

Sample preparation

C. elegans strains were cultured as described previously[64]. All strains were grown at either 15°C or 25°C.

C. elegans wild-type strain N2 (Bristol) and the *let-7(n2853)* mutant strain MT7626 were grown on 9-cm nematode growth medium (NGM) agar plates seeded with a lawn of the *E. coli* strain OP50. N2 and *let-7(n2853)* worms were always grown in parallel (3 biological replicates total). Protein extracts were generated from synchronized late L4 larval stage animals (before vulval bursting) which were grown at 25°C. Worms were harvested and washed three times in M9 media. Generation of the protein extract has been described previously[37]. The buffer used was 50 mM Tris/HCl (pH 8.3), 5 mM EDTA, 8 M urea and 0.125% SDS. Cell debris were removed by sequential centrifugation (4000 g for 5 min followed by 16 000 g for 5 min) and the protein concentrations of the purified extracts were determined by using the Bradford reagent (Sigma-Aldrich). The protein concentrations of N2 and *let-7(n2853)* extracts were adjusted to each other in order to minimize biases for the subsequent Isotope Coded Affinity Tag (ICAT, Applied Biosystems) labeling[33].

ICAT labeling, tryptic digestion of the samples, and the isolation and clean up of ICAT labeled cysteine containing peptides were performed as described in Shiio and Aebersold (2006)[65]. N2 extracts were always labeled with the heavy ICAT reagent and *let-7(n2853)* extracts with the light ICAT reagent. A total of 5 mg per sample and replicate was labeled, resulting in approximately 500 µg ICAT-labeled peptides. The peptide mixtures were cleaned by Sep-Pak tC18 cartridges (Waters) and eluted with 60% acetonitrile. The resulting peptide samples were separated according to the isoelectric point of the peptides by off-gel electrophoresis (OGE; see Ref.27) using a pH 3-10 IPG strip

(AmershamBiosciences), and a 3100 OFFGEL Fractionator (AgilentTechnologies) with collection in 24 wells. Prior to electrofocusing, the peptides were evaporated to dryness in a centrifugal vacuum concentrator and solubilized in a separation medium containing 7 M Urea, 2 M thiourea, 1% w/v DTT, 5% v/v glycerol, 1% v/v carrier ampholytes mixture (IPG buffer pH 3.0-10.0, GE Healthcare) loaded in all wells and the potential was fixed at 8000 V with a maximum current set at 50 µA. Peptides collected in each well were cleaned by Sep-Pak tC18 cartridges (Waters) and eluted with 60% acetonitrile. All peptide samples were evaporated in a vacuum centrifuge to dryness, resolubilized in 2% acetonitrile and 0.1% formic acid and frozen at -20°C until analysed on the mass spectrometer.

RT-qPCR

Prior to protein isolation, a small aliquot of intact animals of each biological replicate (three times N2 wild-type animals and three times *let-7(n2853)* animals – see "sample preparation" above) was frozen, and subsequently used for total RNA isolation. Total RNA was isolated using the Nucleo Spin RNA II kit (Marcherey-Nagel) according to the manufacturer's instructions. After total RNA isolation, genomic DNA was further digested by DNase I using the Turbo DNA-free kit (Ambion) according to the manufacturer's instructions. Total RNA concentrations were determined with the Nanodrop device (Thermo Fisher Scientific). RNA reverse transcription (RT) was performed using the Transcriptor High Fidelity cDNA Synthesis Kit (Roche) with oligo-(dT) primers, according to the manufacturer's recommendations using equal amounts of RNA (4 times 2 µg) for each sample. qPCR reactions were performed in technical duplicate for each of the biological triplicates using MESA Green qPCR Mastermix Plus for SYBR Assay (Eurogentec), according to the manufacturer's recommendations, on an ABI 7900 HT Sequence Detection System coupled to ABI Prism 7900 SDS 2.2 Software (Applied Biosystems). Relative transcript levels were calculated using the $2^{-\Delta Ct}$ method[66]. Most primer pairs were designed via the Roche Universal Probe Library. All the primer pairs used are listed in Supplemental Table 8.

Polysomal profile analysis and subsequent RT-qPCR

The polysomal profile analysis and subsequent RT-qPCR was performed using the same polysomal fractions and protocols as Ding XC and Grosshans H (2009)[49]. The experiments were performed in triplicate.

We could not develop an RT-qPCR assay specific for *let-526b* only, as there is just a small region (< 50bp) in this splice form that is not present in *let-526a*. Instead we used primers specific for both splice forms.

The primers used for *let-526a* specifically were the following:

Fwd:accacgaccaccatatccat

Rev:cgggcattgtagaagagagc

The primers for both *let-526a* and *let-526b* were:

Fwd:tcgccgagagattactcgtt

Rev:agaagcgatgcaaagagcat

RNAi experiments

Gene knockdown was achieved through RNAi by feeding as published[67,68,69,70,17]. Media supplement were used at the following concentrations: ampicillin, 200 µg/ml; isopropyl-β-D-thiogalactopyranoside (IPTG), 2 mM. All the experiments were performed at 25°C. About 100-150 synchronized L1s were placed on IPTG/AMP NGM agarose plates seeded with 200 µl *E. coli* expressing double-stranded RNA (dsRNA). The worms were scored 72 hours later (adult stage) for suppression of lethality. Clones were regarded as positive when at least 20% of the animals were viable as adults. All the clones used were verified by sequencing for their correct insert. All RNAi plasmids used are listed in Supplemental Table 9.

Lethality assays for *C. elegans* mutant strains

All the experiments were performed at 25°C and in quadruplicate. About 100-150 synchronized L1s were placed on NGM agarose plates seeded with 250 µl OP50 *E. coli* bacteria. The worms were scored 48 hours later (= 12 hours post L4) for suppression of lethality. Following strains were tested: MT7626 (*let-7(n2853)*), FX00600 (*ztf-7(tm600)*), WS5673 (*ztf-7(tm600); let-7(n2853)*). At least 20% of the animals had to be viable in the double mutant animals (WS5673) in order to be regarded as a successful suppressor.

It should be noted that 24 hours post L4 most double mutant animals (WS5673) were dead, therefore suggesting more a lethality delay than a true suppression. A developmental delay in WS5673 animals could be excluded as the survivors at the 12 hours post L4 time point had fully developed gonads with oocytes and at least 60% of the survivors also had embryos.

Cloning of 3' UTRs from candidate genes

pEM393 is a dual luciferase Gateway (Invitrogen) compatible vector, adapted from the psiCHECK-II vector (Promega). The 3'UTRs of F46B6.7 (*ztf-7*), C12C8.3 (*lin-41*), F11A1.3a (*daf-12*), F13D11.2 (*hbl-1*), F36A4.7 (*ama-1*) and T04C12.6 (*act-1*) were cloned directly downstream of the Firefly Luciferase gene. The 3'UTRome *C. elegans* database[71] (utrome.org) and Wormbase (www.wormbase.org) were used to retrieve the sequences for the 3'UTRs of interest. Supplemental Table 10 lists the primers used for the PCR reaction and the length of each putative 3'UTR sequence cloned. Gateway cloning was performed according to the manufacturer's instructions (Invitrogen). Briefly, the sequences of interest were amplified using the *attB* adapter primer PCR protocol to generate PCR clones containing the 3'UTR flanked by respective *attB* sites (*attB1* site at the 5' end and the *attB2* site the 3' end). The PCR product was recombined into pDONR221 by the BP reaction to create the entry clone set (see Supplemental Table 10). The entry clones were verified by sequencing and then recombined with the destination vector pEM393 to generate the expression clone set via the LR reaction (see Supplemental Table 10). The expression clones were again verified by sequencing and used for the subsequent luciferase assays.

Luciferase assay

The reactions were performed in 96 well plates. miRNA mimics or inhibitors were ordered from Dharmacon. 150 ng of the dual luciferase expression clone containing the 3'UTR of interest and 10 pmol of the either the human let-7a mimic, the control mimic (*C. elegans* miR-67), the human let-7a inhibitor or the control inhibitor (against *C. elegans* miR-67) were transfected into HeLa cells (10 000 cells per reaction) in triplicate. The Dual-Glo Luciferase assay system (Promega) was used two days

post-transfection, according to the manufacturer's instructions. All the firefly luciferase readouts were first normalized to their matching renilla luciferase readouts. Those readouts were further normalized to empty vector (pEM393 vector without any 3'UTR) controls and then the let-7a readouts (mimics and inhibitors) were normalized to their respective oligo controls.

Design of SRM assays

861 genes of interest were selected based on literature, computational prediction algorithms, experimental evidence and MS detectability (random control). PTPs were selected based on a large shotgun proteomics data set[37]. This *C. elegans* proteome atlas data set was filtered for a peptide-spectrum match false discovery rate of 0.17% corresponding to a protein identification false discovery rate of 5%[38]. Proteotypic peptides needed to contain at least one cysteine[33] and doubly charged peptides with a high number of identifications were preferred. For each peptide, 4 to 8 fragment ions from the y-ion series were computed. Fragment ions (Q3) with an m/z above the peptide ion (Q1) and with a defined minimal distance to the peptide ion were chosen ($m/z_{Q3} - m/z_{Q1} \geq 50$ Thomson). The peptide ion/fragment ion (Q1/ Q3) transitions were used to trigger the acquisition of triple quadrupole (QQQ) MS/MS spectra of the peptides of interest in *C. elegans* whole worm extracts and in off gel electrophoresis (OGE) fractionated samples. Proteotypic peptides for additional 19 proteins not contained in the *C. elegans* proteome atlas were found using SRM triggered MS/MS. For the samples derived from the OGE fractionations, the isoelectric points of the peptides were predicted using BioPerl[72] and peptides were targeted in the predicted fraction and in the two neighboring fractions if available.

Database search and extraction of optimal SRM transitions

The data was converted from the raw .wiff to the .mzXML format using the program mzWiff (version 3.5.3, build Apr 16 2008 14:40:24). The MS/MS spectra from the SRM triggered MS/MS experiments were searched against wormpep140 (www.wormbase.org) using Sequest on a Sorcerer machine (Sorcerer™-SEQUEST®, 3.10.4 release) with light ICAT as static modification and heavy ICAT and/or. oxidized methionine as variable

modifications. Precursor mass tolerance was set to 1.5 Da and the data were searched fully tryptic with maximal two missed cleavages. The data was filtered with a peptide-spectrum match FDR of 2.5% using PeptideProphet[73]. Three transitions for each proteotypic peptide were generated by extracting the three highest fragment ions and the retention time of the peptide from the triple quadrupole MS/MS. All transitions used for quantification in this study are listed in Supplemental Table 7.

Mass spectrometry analysis

All peptide samples were analyzed on a hybrid triple quadrupole/ion trap mass spectrometer (4000QTrap, ABI/MDS-Sciex) equipped with a nanoelectrospray ion source. Chromatographic separations of peptides were performed on a Tempo nano LC system (Applied Biosystems) coupled to a 16 cm fused silica emitter, 75 μm diameter, packed with a Magic C18 AQ 5 μm resin (Michrom BioResources). Peptides were loaded onto a trapping column from a cooled (4°C) Tempo autosampler and separated with a linear gradient of acetonitrile/water, containing 0.1% formic acid, at a flow rate of 300 nl/min. A gradient from 5 to 30% acetonitrile in 30 or 60 minutes was used. Collision energies used for both SRM and MS/MS analyses were calculated according to the formulas: CE = 0.044 * m/z + 5.5 or CE = 0.051 * m/z + 0.5 (CE, collision energy, m/z, mass-to-charge ratio of the precursor ion) for doubly and triply charged precursor ions, respectively (see Ref.27).

Validation: In the SRM assays validation phase, the mass spectrometer was operated in multiple reaction monitoring mode, triggering acquisition of a full MS/MS spectrum upon detection of an SRM trace (MRM-triggered MS/MS, threshold 200 ion counts). The set of SRM transitions generated as previously described was split into multiple MS-methods and analyzed in several runs. Each SRM acquisition was performed with Q1 and Q3 operated at unit resolution (0.7 m/z half maximum peak width). An average of 60 transitions (dwell time 20 ms/transition) per run was used for the measurements. MS/MS spectra were acquired in enhanced product ion (EPI) mode for the two highest SRM transitions, using dynamic fill time, Q1 resolution low, scan speed 4000 amu/s, m/z range 300-1400.

Quantification: An average of 60 transitions per run was used for the measurements. The quantification measurements were done in the scheduled SRM mode (retention time window: 900 seconds; target scan time: 2 seconds)

Quantitative and statistical analysis

Peak height for the transitions associated to the *let-7(n2853)* (light ICAT label) and wild-type (heavy ICAT label) derived peptides were quantified using the software MultiQuant v. 1.1 Beta (Applied Biosystems). Log_2 fold changes were calculated for each transition separately. These values were then normalized using 11 proteotypic peptides (see Supplemental Figure 1a and Supplemental Figure 2) on each biological replicate separately. To test for statistically significant abundance changes, a two sided one sample t-test was done on the normalized log_2 fold changes of the transitions grouped according to protein (μ equal to zero). To generate our list of regulated candidates we used a *P*-value ≤ 0.01 cutoff.

REFERENCES

1. Bartel, D.P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215-33 (2009).
2. Ventura, A. & Jacks, T. MicroRNAs and cancer: short RNAs go a long way. *Cell* **136**, 586-91 (2009).
3. Stark, A., Brennecke, J., Bushati, N., Russell, R. & Cohen, S. Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3′UTR Evolution. *Cell* **123**, 1133–1146 (2005).
4. Lall, S. et al. A Genome-Wide Map of Conserved MicroRNA Targets in C. elegans. *Current Biology* **16**, 460–471 (2006).
5. Griffiths-Jones, S., Saini, H.K., Dongen, S.V. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Research* **36**, D154–D158 (2007).
6. Lim, L.P. et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769-73 (2005).
7. Ziegelbauer, J.M., Sullivan, C.S. & Ganem, D. Tandem array–based expression screens identify host mRNA targets of virus-encoded microRNAs. *Nat Genet* **41**, 130–134 (2009).
8. Zhang, L. et al. Systematic Identification of C. elegans miRISC Proteins, miRNAs, and mRNA Targets by Their Interactions with GW182 Proteins AIN-1 and AIN-2. *Molecular Cell* **28**, 598–613 (2007).
9. Easow, G., Teleman, A.A. & Cohen, S.M. Isolation of microRNA targets by miRNP immunopurification. *RNA* **13**, 1198–1204 (2007).
10. Beitzinger, M., Peters, L., Zhu, J.Y., Kremmer, E. & Meister, G. Identification of human microRNA targets from isolated argonaute protein complexes. *RNA Biol* **4**, 76-84 (2007).
11. Karginov, F.V. et al. A biochemical approach to identifying microRNA targets. *Proc Natl Acad Sci U S A* **104**, 19291-6 (2007).
12. Hendrickson, D.G. et al. Systematic Identification of mRNAs Recruited to Argonaute 2 by Specific microRNAs and Corresponding Changes in Transcript Abundance. *PLoS ONE* **3**, e2126 (2008).
13. Landthaler, M. et al. Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA* **14**, 2580–2596 (2008).
14. Ørom, U.A., Nielsen, F.C. & Lund, A.H. MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Mol Cell* **30**, 460-71 (2008).
15. Lee, R.C., Feinbaum, R.L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-54 (1993).
16. Reinhart, B.J. et al. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**, 901-6 (2000).
17. Grosshans, H., Johnson, T., Reinert, K., Gerstein, M. & Slack, F. The Temporal Patterning MicroRNA Regulates Several Transcription Factors at the Larval to Adult Transition in. *Developmental Cell* **8**, 321–330 (2005).

18. Baek, D. et al. The impact of microRNAs on protein output. *Nature* **455**, 64–71 (2008).
19. Selbach, M. et al. Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008).
20. Nakahara, K. et al. Targets of microRNA regulation in the Drosophila oocyte proteome. *Proc Natl Acad Sci U S A* **102**, 12023-8 (2005).
21. Tian, Z., Greene, A.S., Pietrusz, J.L., Matus, I.R. & Liang, M. MicroRNA-target pairs in the rat kidney identified by microRNA microarray, proteomic, and bioinformatic analysis. *Genome Research* **18**, 404–411 (2008).
22. Vinther, J., Hedegaard, M.M., Gardner, P.P., Andersen, J.S. & Arctander, P. Identification of miRNA targets with stable isotope labeling by amino acids in cell culture. *Nucleic Acids Research* **34**, e107–e107 (2006).
23. Yang, Y., Chaerkady, R., Beer, M.A., Mendell, J.T. & Pandey, A. Identification of miR-21 targets in breast cancer cells using a quantitative proteomic approach. *Proteomics* **9**, 1374-84 (2009).
24. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* **4**, 222 (2008).
25. Stahl-Zeng, J. et al. High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. *Mol Cell Proteomics* **6**, 1809-17 (2007).
26. Lange, V. et al. Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring. *Mol Cell Proteomics* **7**, 1489-500 (2008).
27. Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B. & Aebersold, R. Full Dynamic Range Proteome Analysis of S. cerevisiae by Targeted Proteomics. *Cell* (2009).doi:10.1016/j.cell.2009.05.051
28. Picotti, P. et al. A database of mass spectrometric assays for the yeast proteome. *Nat Methods* **5**, 913-4 (2008).
29. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol* **6**, 577-83 (2005).
30. Mallick, P. et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* **25**, 125–131 (2007).
31. Fusaro, V.A., Mani, D.R., Mesirov, J.P. & Carr, S.A. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat Biotechnol* **27**, 190-8 (2009).
32. Schmidt, A., Kellermann, J. & Lottspeich, F. A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* **5**, 4–15 (2005).
33. Gygi, S.P. et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**, 994-9 (1999).
34. Krijgsveld, J. et al. Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics. *Nat Biotechnol* **21**, 927–931 (2003).
35. Ong, S. et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics* **1**, 376-386 (2002).
36. Gstaiger, M. & Aebersold, R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat. Rev. Genet* **10**, 617-627 (2009).
37. Schrimpf, S.P. et al. Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes. *PLoS Biol* **7**, e48 (2009).

38. Reiter, L. et al. Protein identification false discovery rates for very large proteomics datasets generated by tandem mass spectrometry. *Mol. Cell Proteomics* (2009).doi:10.1074/mcp.M900317-MCP200
39. Pasquinelli, A.E. et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86-9 (2000).
40. Johnson, S.M. et al. RAS is regulated by the let-7 microRNA family. *Cell* **120**, 635-47 (2005).
41. Ding, X.C., Slack, F.J. & Grosshans, H. The let-7 microRNA interfaces extensively with the translation machinery to regulate cell differentiation. *Cell Cycle* **7**, 3083-90 (2008).
42. Abrahante, J.E. et al. The Caenorhabditis elegans hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. *Dev Cell* **4**, 625-37 (2003).
43. Slack, F.J. et al. The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. *Mol Cell* **5**, 659-69 (2000).
44. Lin, S. et al. The C elegans hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. *Dev Cell* **4**, 639-50 (2003).
45. Watanabe, Y. et al. Computational analysis of microRNA targets in Caenorhabditis elegans. *Gene* **365**, 2-10 (2006).
46. Papadopoulos, G.L., Reczko, M., Simossis, V.A., Sethupathy, P. & Hatzigeorgiou, A.G. The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res* **37**, D155-8 (2009).
47. Xiao, F. et al. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* **37**, D105-10 (2009).
48. Vasudevan, S., Tong, Y. & Steitz, J.A. Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318**, 1931-4 (2007).
49. Ding, X.C. & Grosshans, H. Repression of C. elegans microRNA targets at the initiation level of translation requires GW182 proteins. *EMBO J* **28**, 213-22 (2009).
50. Ding, L., Spencer, A., Morita, K. & Han, M. The Developmental Timing Regulator AIN-1 Interacts with miRISCs and May Target the Argonaute Protein ALG-1 to Cytoplasmic P Bodies in. *Molecular Cell* **19**, 437–447 (2005).
51. Ruby, J. et al. Large-Scale Sequencing Reveals 21U-RNAs and Additional MicroRNAs and Endogenous siRNAs in C. elegans. *Cell* **127**, 1193–1207 (2006).
52. Nottrott, S., Simard, M.J. & Richter, J.D. Human let-7a miRNA blocks protein production on actively translating polyribosomes. *Nat Struct Mol Biol* **13**, 1108–1114 (2006).
53. Andachi, Y. A novel biochemical method to identify target genes of individual microRNAs: Identification of a new Caenorhabditis elegans let-7 target. *RNA* **14**, 2440–2451 (2008).
54. Picotti, P. et al. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat. Methods* (2009).doi:10.1038/nmeth.1408
55. Addona, T.A. et al. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol* (2009).doi:10.1038/nbt.1546
56. Johnson, S.M., Lin, S.Y. & Slack, F.J. The time of appearance of the C. elegans let-7

microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. *Dev Biol* **259**, 364-79 (2003).
57. Prakash, A. et al. Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. *J. Proteome Res* **8**, 2733-2739 (2009).
58. Grosshans, H. & Filipowicz, W. Proteomics Joins the Search for MicroRNA Targets. *Cell* **134**, 560–562 (2008).
59. Boyerinas, B. et al. Identification of let-7-regulated oncofetal genes. *Cancer Res* **68**, 2587-91 (2008).
60. Brunner, E. et al. A high-quality catalog of the Drosophila melanogaster proteome. *Nat Biotechnol* **25**, 576–583 (2007).
61. Loevenich, S.N. et al. The Drosophila melanogaster PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation. *BMC Bioinformatics* **10**, 59 (2009).
62. Desiere, F. et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol* **6**, R9 (2005).
63. Baerenfaller, K. et al. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* **320**, 938-41 (2008).
64. Brenner, S. The genetics of Caenorhabditis elegans. *Genetics* **77**, 71-94 (1974).
65. Shiio, Y. & Aebersold, R. Quantitative proteome analysis using isotope-coded affinity tags and mass spectrometry. *Nat Protoc* **1**, 139–145 (2006).
66. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-8 (2001).
67. Fraser, A.G. et al. Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature* **408**, 325-30 (2000).
68. Kamath, R.S. et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231-7 (2003).
69. Timmons, L. & Fire, A. Specific interference by ingested dsRNA. *Nature* **395**, 854 (1998).
70. Rual, J. et al. Toward improving Caenorhabditis elegans phenome mapping with an ORFeome-based RNAi library. *Genome Res* **14**, 2162-8 (2004).
71. Mangone, M., Macmenamin, P., Zegar, C., Piano, F. & Gunsalus, K.C. UTRome.org: a platform for 3'UTR biology in C. elegans. *Nucleic Acids Res* **36**, D57-62 (2008).
72. Stajich, J.E. et al. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**, 1611-8 (2002).
73. Keller, A., Eng, J., Zhang, N., Li, X. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **1**, 2005.0017 (2005).

Figure 1: Jovanovic M, Reiter L *et al.*

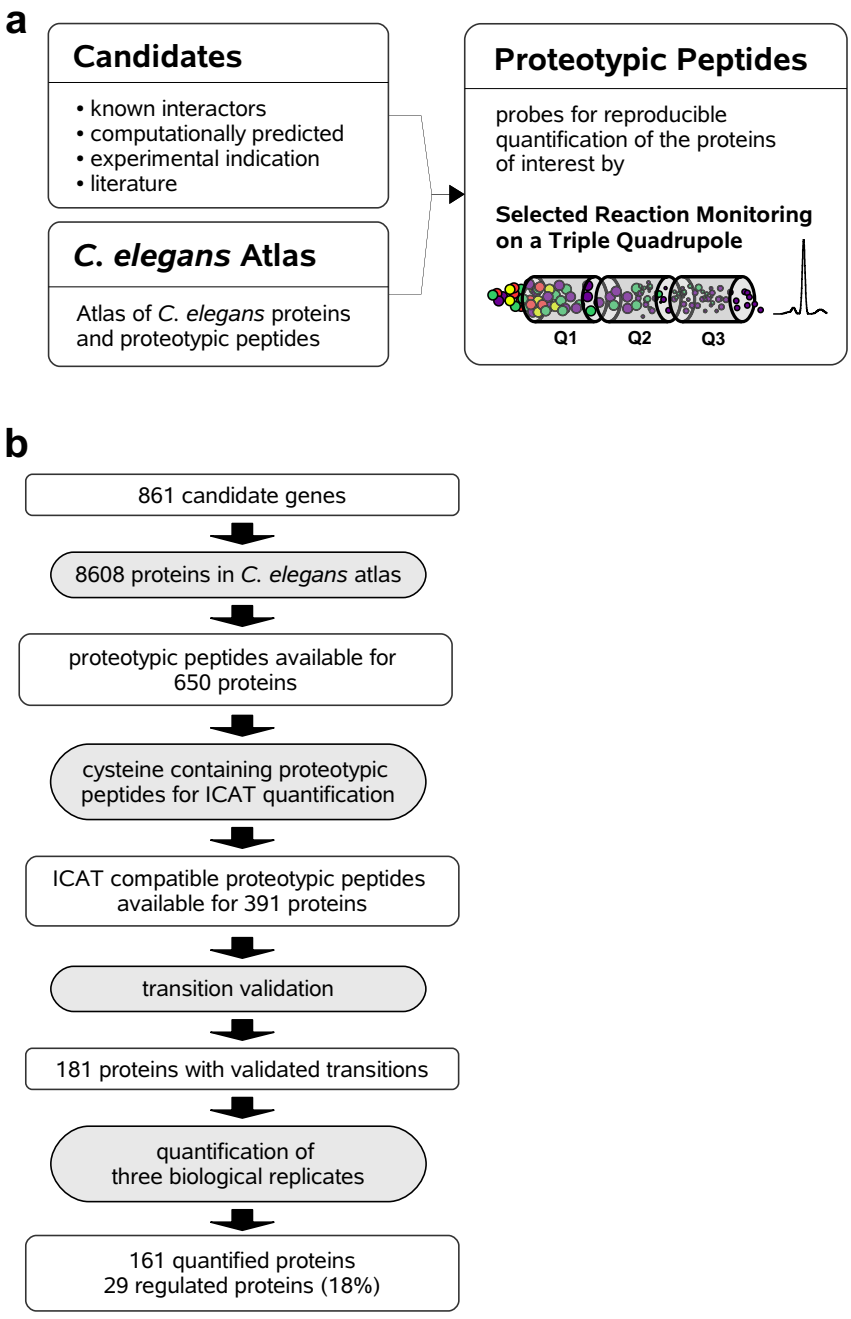


Figure 2: Jovanovic M, Reiter L *et al.*

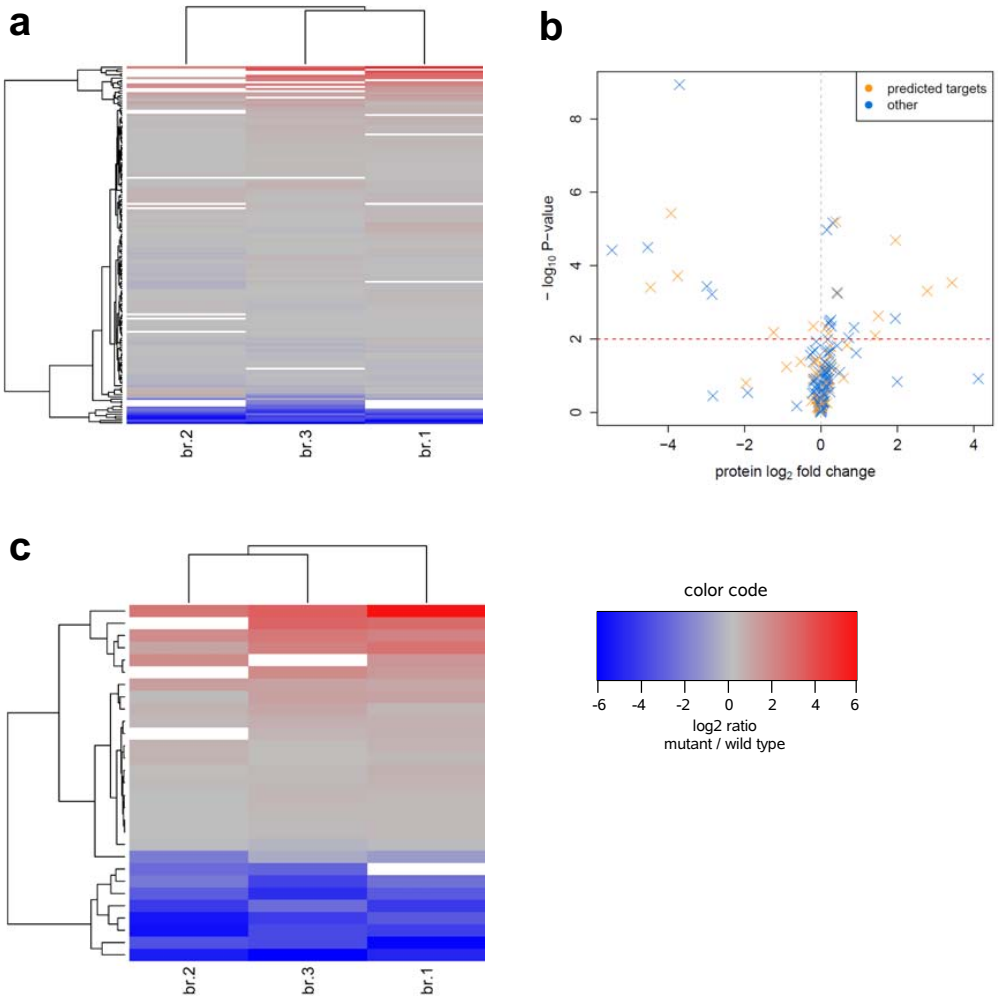


Figure 3: Jovanovic M, Reiter L *et al.*

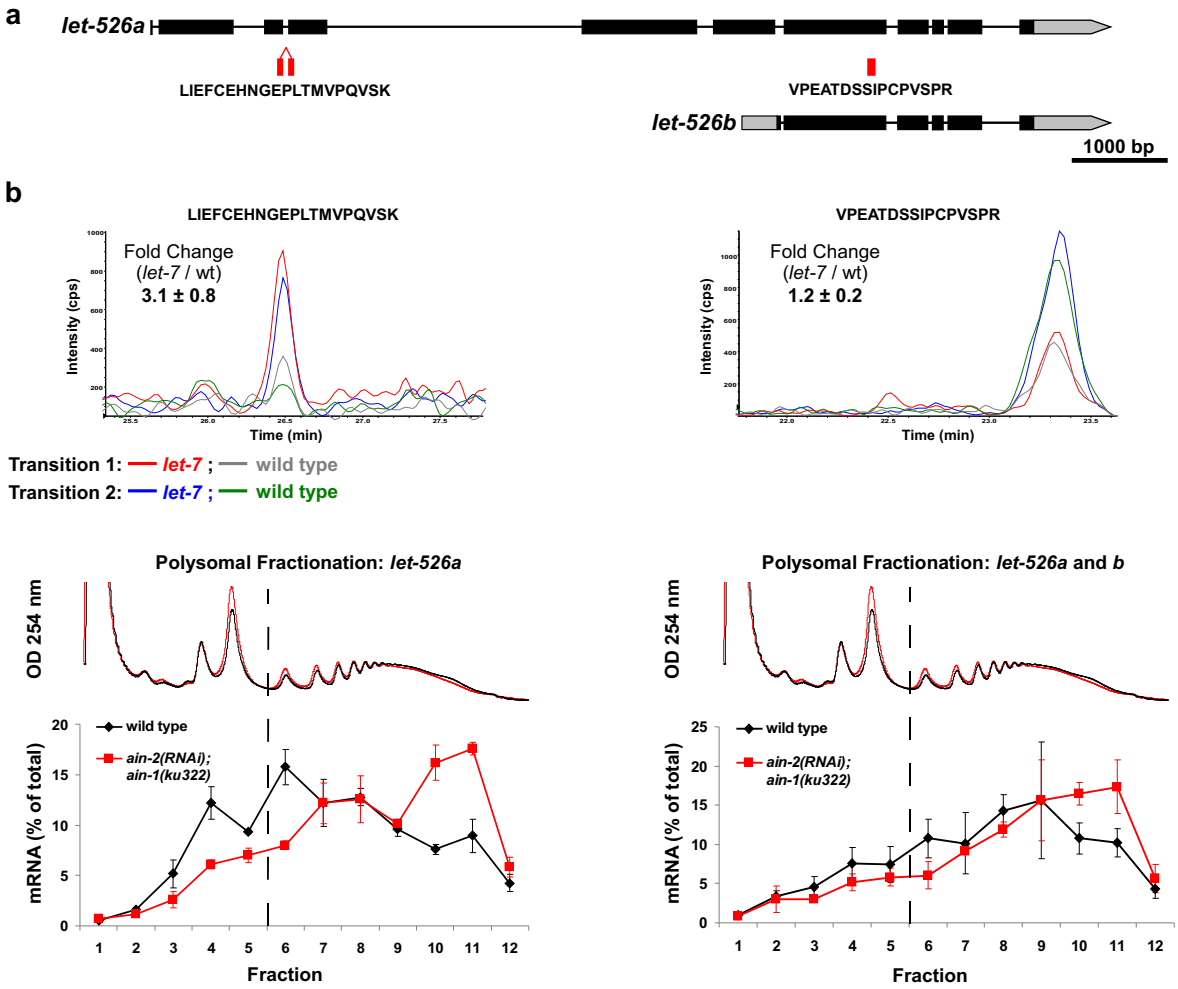
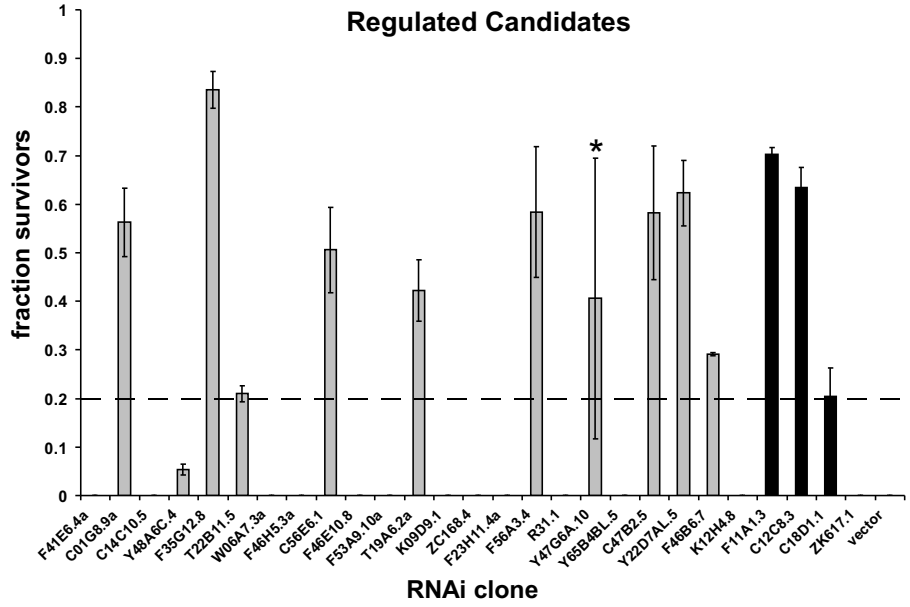
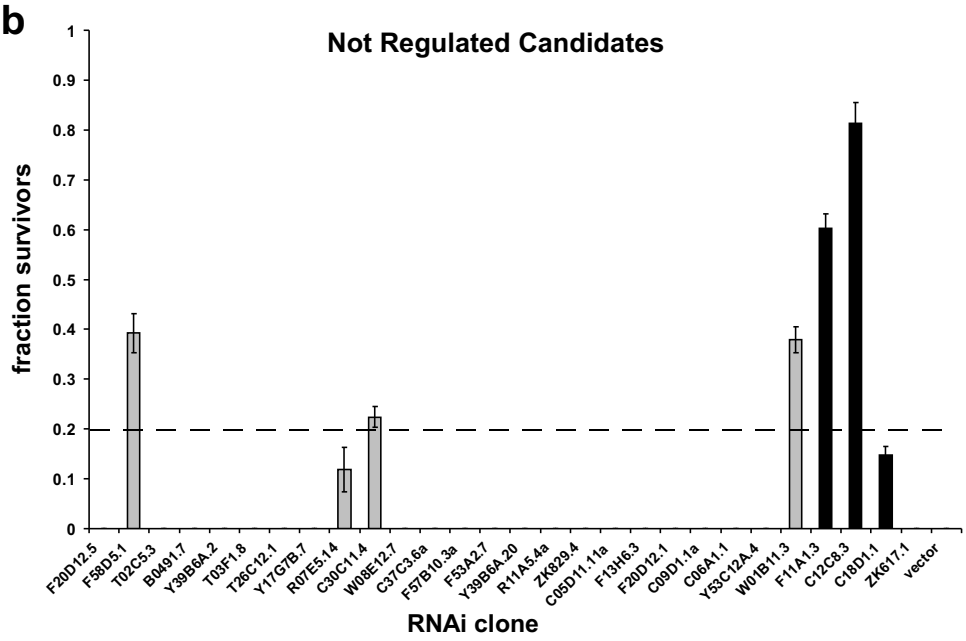


Figure 4: Jovanovic M, Reiter L *et al.*

a



b



c

fraction survivors cutoff	regulated	not regulated	*P*-value
0.2	10	3	0.024
0.4	8	0	0.002

Figure 5: Jovanovic M, Reiter L *et al.*

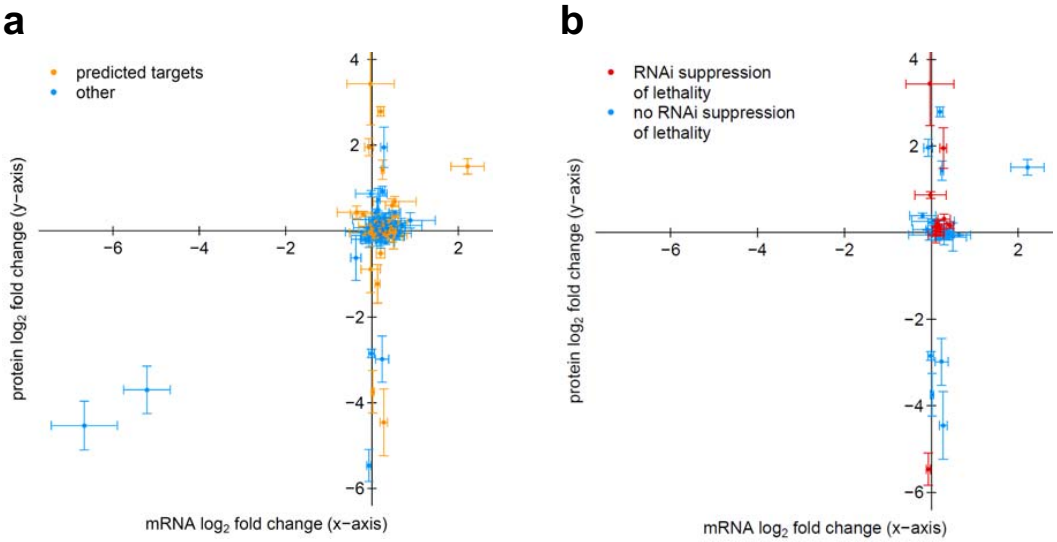
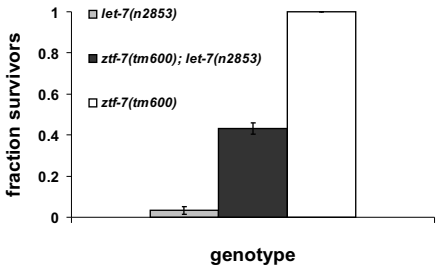


Figure 6: Jovanovic M, Reiter L *et al.*

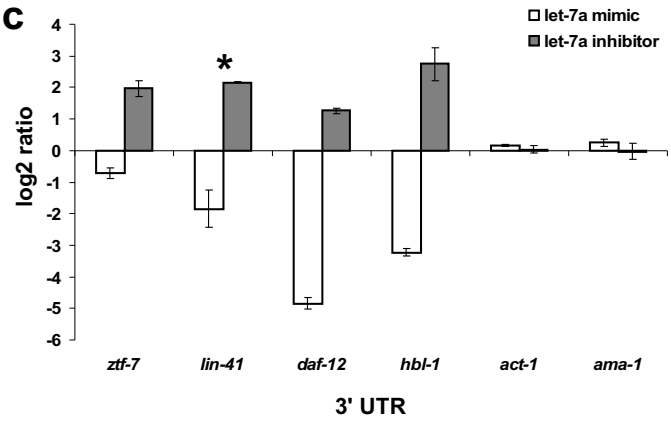
a

Gene	average Fold Change (*let-7* / wt)	*P*-value
ztf-7	1.1	0.0048

b



c



7.2 *Arabidopsis* female gametophyte gene expression map reveals similarities between plant and animal gametes

7.2.1 Preface

My contribution to this paper consisted of the GO Term enrichment analysis when comparing human and *Arabidopsis* egg cells, including the orthology computation for these two species.

Current Biology 20, 506–512, March 23, 2010 ©2010 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2010.01.051

Report

Arabidopsis Female Gametophyte Gene Expression Map Reveals Similarities between Plant and Animal Gametes

Samuel E. Wuest,[1,2] Kitty Vijverberg,[1,8] Anja Schmidt,[1] Manuel Weiss,[3,4,5,6] Jacqueline Gheyselinck,[1,9] Miriam Lohr,[7] Frank Wellmer,[2] Jörg Rahnenführer,[7] Christian von Mering,[3,6] and Ueli Grossniklaus[1,4,*]

[1]Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, 8008 Zürich, Switzerland

[2]Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland

[3]Institute of Molecular Life Sciences, University of Zürich

[4]Center for Model Organism Proteomes

[5]PhD Program in Molecular Life Sciences

[6]Swiss Institute of Bioinformatics Winterthurerstrasse 190, 8057 Zürich, Switzerland

[7]Fakultät Statistik, Technische Universität Dortmund, Vogelpothsweg 87, 44227 Dortmund, Germany

Summary

The development of multicellular organisms is controlled by differential gene expression whereby cells adopt distinct fates. A spatially resolved view of gene expression allows the elucidation of transcriptional networks that are linked to cellular identity and function. The haploid female gametophyte of flowering plants is a highly reduced organism: at maturity, it often consists of as few as three cell types derived from a common precursor [1, 2]. However, because of its inaccessibility and small size, we know little about the molecular basis of cell specification and differentiation in the female gametophyte. Here we report expression profiles of all cell types in the mature *Arabidopsis* female gametophyte. Differentially expressed posttranscriptional regulatory modules and metabolic pathways characterize the distinct cell types. Several transcription factor families are overrepresented in the female gametophyte in comparison to other plant tissues, e.g., type I MADS domain, RWP-RK, and reproductive meristem transcription factors. PAZ/Piwi-domain encoding genes are upregulated in the egg, indicating a role of epigenetic regulation through small RNA pathways—a feature paralleled in the germline of animals [3]. A comparison of human and *Arabidopsis* egg cells for enrichment of functional groups identified several similarities that may represent a consequence of coevolution or ancestral gametic features.

Results and Discussion

The plant life cycle alternates between a diploid sporophyte and a haploid gametophyte generation. During evolution, the gametophyte generation has been reduced in size and complexity [1, 2]. Because of its simple structure, the female gametophyte of flowering plants is an ideal system to determine a complete expression map of an organism. However, its small number of cells and inaccessibility have made molecular and genome-wide studies difficult. To determine cell-type-specific expression profiles in the female gametophyte of *Arabidopsis*, we combined laser-assisted microdissection (LAM) of individual cells with the Affymetrix ATH1 GeneChip, a microarray platform commonly used in *Arabidopsis* research. LAM allowed us to dissect the cells of the mature female gametophyte with little cross-contamination (Figures 1A–1C; see Figure S1 available online). RNA isolated from 300 to 800 cells per sample was amplified via a linear amplification protocol and hybridized to ATH1 GeneChips (Table S1). Cell-type-specific transcriptomes were obtained for the synergids and the two female gametes, the egg and central cell (Figures 1D–1G; Figures S1A–S1N).

A consequence of linear amplification in combination with small input amounts of RNA is the predominant amplification of 3′ mRNA ends, resulting in a loss of signal at the 5′ mRNA end. The default algorithm generally used to test whether a gene is detectable above background levels performs poorly on data from amplified samples [4]. Therefore, we applied a novel algorithm to create present/absent calls, hereafter denoted *At*PANP (Supplemental Experimental Procedures). We assessed *At*PANP with measures of precision, such as overlaps between biological replicates and accuracy, i.e., by checking its predictive power for genes known to be expressed in the female gametophyte (69 genes; Table S2). Our new algorithm outperforms the default method on our data set (Figures S1I–S1R). Using this robust statistical method, we estimate the mature female gametophyte to express about 8,850 of the 20,777 genes present on the array (conservative estimate; Table S1). This is slightly lower than our conservative estimate of 9,220 genes expressed in pollen (male gametophyte) and sperm [5]. Because of complexity reduction during amplification, we may slightly underestimate the real transcriptome size such that mature male and female gametophytes have similar transcriptional activities.

To validate the microarray data, we used alternative approaches (Table S2; Supplemental Results) such as (1) in situ hybridization (Figures 2A–2C), (2) analysis of putative *cis*-regulatory elements driving the *GUS* reporter gene (Figures 2D–2F; Figures S2A and S2B), (3) characterization of gene and enhancer traps (Figures S2C and S2D), (4) comparison to published data (Figure 2G; Table S2), and (5) comparison to maize egg cell EST data (Figures S2E and S2F). Based on these extensive validations, we conclude that the data set reported here is accurate and can be used to predict preferential expression of genes in the female gametophyte at the level of its specific cell types.

Female gametophytic cells are closely related with regard to their cell lineage yet play distinct roles during reproduction [1, 2]. We found 1345 differentially expressed genes at a low-stringency cutoff of an unadjusted analysis of variance p value below 0.01, and 431 at a false discovery rate below 0.05. The majority of these were enriched in only one cell type, as shown by either subgrouping through pairwise t tests or hierarchical

*Correspondence: grossnik@botinst.uzh.ch

[8]Present address: Keygene N.V., P.O. Box 216, 6700 AE Wageningen, The Netherlands

[9]Present address: Department of Plant Molecular Biology, University of Lausanne, 1015 Lausanne, Switzerland

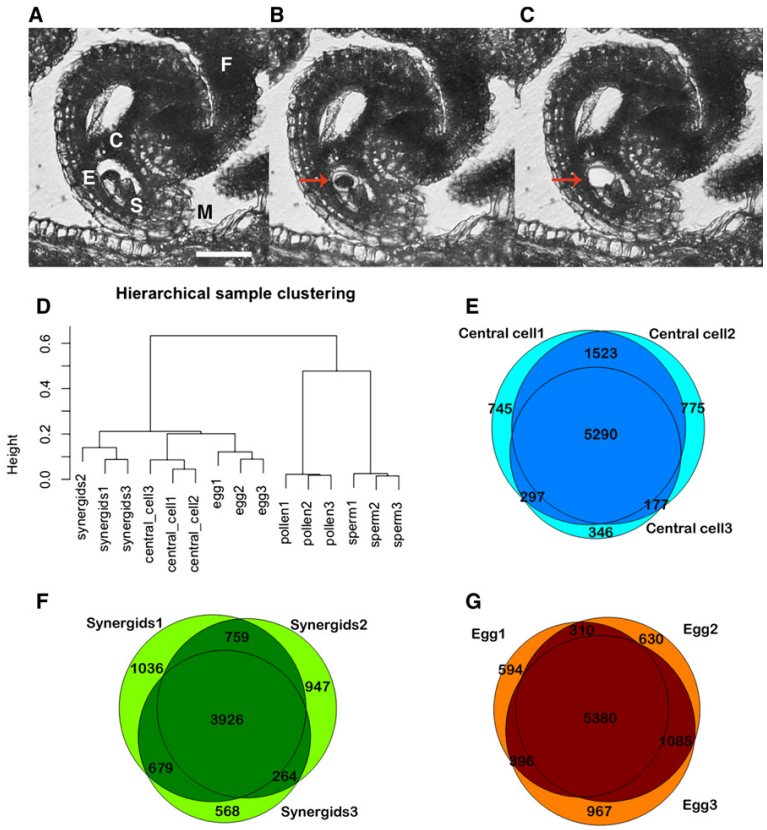


Figure 1. Laser-Assisted Microdissection and Subsequent Analysis of Transcriptomes from Populations of Individual Female Gametophytic Cell Types

(A) Dissection of the egg cell from a mature embryo sac; 8 μm section through an ovule bearing a mature embryo sac before laser microdissection with the MMI SL μCut instrument. The following abbreviations are used: M, micropyle; S, synergids; E, egg cell; C, central cell; F, funiculus. Scale bar represents 37 μm.

(B) The ultraviolet irradiation laser beam has been applied in order to isolate the egg cell (arrow). The laser cut has a diameter of 1–2 μm.

(C) The egg cell has been removed with an MMI isolation cap. After the isolation of the egg, the two remaining cell types (central cell and synergids) were collected on separate isolation caps (see also Figure S1).

(D) Hierarchical agglomerative sample clustering (correlation distance) of male and female gametophytic cell types; note that biological replicates are grouped together, demonstrating that the data are reproducible. Arrays from female gametophytic cell types form a close cluster when compared to male gametophytic cells [5].

(E–G) Overlaps of predictions of gene expression (present calls) when determined by a novel, empirical approach (*At*PANP). The algorithm determines whether a gene is expressed on an array by comparing its signal against a background distribution calculated by the use of negative probes. The Venn diagrams show present call overlaps in the three biological replicates for *At*PANP present calls (p value cutoff = 0.02; see also Table S1). Genes whose present call p values were below the cutoff in at least two of three replicates were considered present in a given cell type (darker areas): egg cell, 7171 genes; central cell, 7287 genes; synergids, 5628 genes. (E) versus (F) versus (G) are not to scale (see also Figure S1).

agglomerative clustering (Figure 3A; Table S3). This agrees with a recent study on the expression of 43 genes in the female gametophyte, of which 41 were strongly enriched in one cell type [6]. A functional gene classification and higher-level analysis suggests that the cells of the mature female gametophyte exhibit differential gene expression in distinct posttranscriptional and epigenetic regulatory mechanisms and metabolic pathways (Figure 3B). Recently, it was shown that auxin patterns the female gametophyte [7], but how this positional information controls cell specification is unknown. Our higher-level analysis did not suggest cell-type-specific differences in auxin readout. However, we identified candidates that could function in developmental events triggered by auxin: the auxin response factor *ARF17* and the polar auxin transport regulator *MKK7* genes exhibited elevated levels in egg and central cell, respectively, whereas the auxin-responsive gene *AT2G16580* was enriched in the entire gametophyte (Figure S3A; Table S3).

A dominant feature of egg cells is the relatively high expression of genes encoding the double-stranded RNA-binding factors DCL1, HYL1, and AT4G00420, a paralog of RNASE THREE-LIKE PROTEIN 1, in addition to RISC components such as AGO1. PAZ and Piwi domain-encoding genes are highly enriched among differentially expressed genes with predominant expression in the egg (Figure S3B; Table S3). In contrast, *SGS3*, involved in various gene silencing pathways, shows elevated expression in the central cell and is possibly involved in small interfering RNA (siRNA) production [8]. This suggests an important role of RNA-based silencing mechanisms in the female gametes, and the large diversity of recently discovered maternal siRNAs in developing *Arabidopsis* seeds [9] may, in part, be explained by maternal deposition of siRNAs in the female gametophyte.

In order to relate the female gametophyte transcriptome to other plant tissues, we compared it with data from 59 different tissues of the *Arabidopsis* sporophyte and male gametophyte (Table S1; Supplemental Experimental Procedures) [4, 5, 10, 11]. First we used sample clustering on binary present/absent calls to assess the overall structure of the female gametophytic cell transcriptomes. They were comparable in size and/or composition to the transcriptomes of male gametophytes or laser-captured embryos but distinct from sporophytic tissues or cell types from root and shoot, which exhibit higher expression activities (Figure 3C). Thus, male and female gametophytes share transcriptome characteristics that are distinct from those of sporophytic tissues. Interestingly, the embryo shares more characteristics with the gametophytes from which it is derived than with the adult sporophyte. Future comparisons with transcriptomes of gametophytes and sporophytes from haploid-dominant plants, such as mosses, may reveal differential gene family expansion or transfer of transcriptional modules from the haploid to the diploid generation [12].

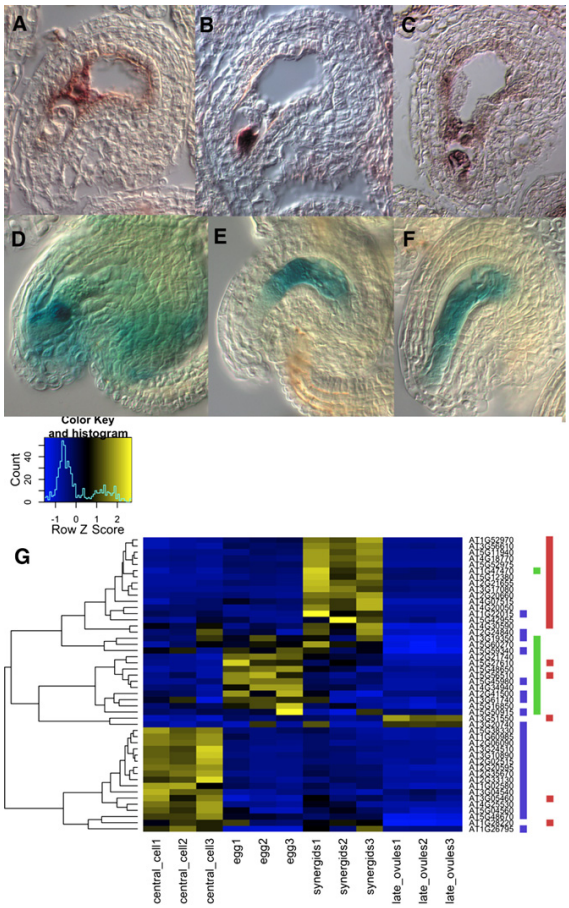


Figure 2. Data Validation by In Situ Hybridization, Promoter GUS Fusions, and Comparison to the Literature

(A–C) In situ hybridization of genes with enriched expression in the female gametophyte.

(A) *AT2G20595*, a gene with unknown function, is highly expressed in the central cell and expressed at low levels in the egg and synergids.

(B) *AT3G17080*, a self-incompatibility-related gene, is highly expressed in the synergids.

(C) The *Arabidopsis* telomerase gene *AT5G16850* shows increased expression in the egg cell and a lower expression level in the central cell.

(D–F) GUS activity in ovules expressing the *GUS* reporter gene under 5′ upstream elements (5′UE) of different genes enriched in the female gametophyte.

(D) The 5′UE of *AT5G48650*, encoding a nuclear transport factor, shows highest activity in the egg.

(E) The 5′UE of *RALF18*, a gene with putative signaling function, is highly expressed in the central cell and at much lower levels in the egg cell and synergids.

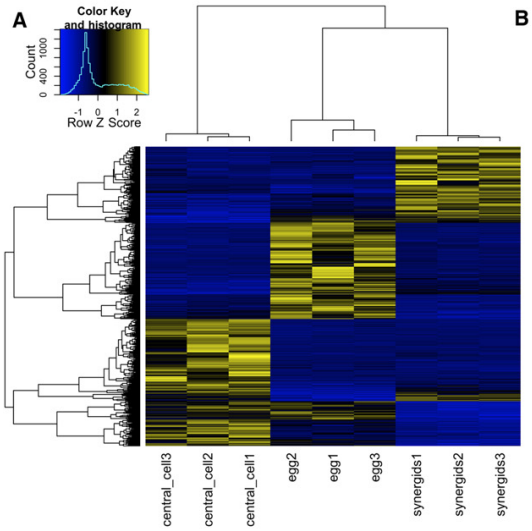
(F) The 5′UE of the MYB64 transcription factor gene *AT5G11050* shows high activity in the whole gametophyte but not in the surrounding sporophytic tissue.

(G) Heat map of expression signals of genes with described differential expression within the female gametophyte. Yellow denotes high expression, blue denotes low expression. Sample/genes were clustered via correlation distance and hierarchical agglomerative clustering, and colors are scaled per row. The color code panels on the right indicate the described preferential expression of a gene according to the literature: preferential expression in synergids, red; egg, green; central cell, blue. Note that apart from <8% disagreement out of 96 contrasts examined, the array data mirrors preferential expression within these cell types (see also Table S2).

The construction of a comprehensive tissue atlas allowed us to identify genes exhibiting enriched expression in female gametophytic cells (Supplemental Experimental Procedures). At stringent conditions, we found 420 genes significantly enriched in one of the cell types (Table S3; Figure S3C), including several genes playing a role in gametophyte development and function: *MYB98* [13] in the synergids and *FIS2* [14], *DME* [15], *CKI1* [16], *UNE6*, and *EDA28* [17] in the central cell. Thus, genes enriched in the three cell types are likely involved in cell-type-specific functions and constitute an important resource for reverse and forward genetic approaches. Recently, small, cysteine-rich defensin-like proteins (DEFLs) were implicated as signaling molecules required for pollen tube guidance in *Zea mays* and *Torenia fournieri* [18, 19]. Of the 33 (of 317) DEFLs [20] present on the array, we found seven highly enriched in the female gametophyte (Figure S3D). Six were predominantly expressed in the central cell but not the synergids, which produce the guidance signal. Whether these DEFLs act as signals remains to be examined, but they might contribute to the recently discovered role of the *Arabidopsis* central cell in pollen tube guidance [21].

We next searched for gene families or groups of genes containing a Pfam domain (Pfam groups) that are globally enriched in female gametophytic cells. Five of the ten gene sets previously found enriched in the female gametophyte [22] were also significantly enriched when examined in the more comprehensive context of our tissue atlas. Seventy-four Pfam groups and 32 gene families were enriched in at least one of the cell types or in the entire female gametophyte ($p < 0.01$; Table S3). Enriched Pfam groups contain a high number of domains of unknown function (DUFs), highlighting the lack of characterization for genes expressed in the female gametophyte (seen: 20 from 74; expected: 7.5; chi-square $p < 0.001$). Gene sets involved in transcriptional, posttranscriptional, and epigenetic regulation, signaling, and cell wall modification were enriched (Figures S3E–S3H; Table S3). Expansins were overrepresented in the transcriptomes of both male and female gametophytes (Figure S3E), as may be expected given their rapid growth, necessitating cell wall biosynthesis. Three groups of transcription factors (TFs) were overrepresented in the whole female gametophyte transcriptome, namely the RWP-RK domain, the MADS domain (predominantly type I), and the reproductive meristem TF families. It was shown that several members of these families are important in sexual plant reproduction [23–28]. Type I MADS-domain TFs were exclusively enriched in reproductive tissues, i.e., male and female gametophytes, and developing embryos (Figure S3F). Of the 28 type I MADS-domain TFs on the array, seven had highest expression in the female gametophyte—including AGL23 [26], AGL61 (DIANA) [27], and AGL80 [24], known to play a functional role in this tissue—whereas AGL62—required for endosperm cellularization [28]—exhibited highest expression in the central cell. Other family members showed highest expression in the male gametophyte (10 genes), embryo (9 genes), or seed (2 genes). These expression patterns suggest a predominant role of type I MADS-domain TFs in sexual reproduction, which may explain their highly dynamic evolutionary history [29]. Other enriched gene families, e.g., those encoding F box or leucine-rich repeat domains (Figures S3G–S3H), have also undergone rapid evolution, possibly correlated with their putative role in reproduction.

Our analysis highlighted that genes encoding PAZ, Piwi, and DUF1785 domains, mainly associated with the *Arabidopsis* Argonaute and Dicer proteins, were globally enriched in the



B

Biol. Process	Term	Signif.	Exp.	P value
GO:0016246	RNA interference	7	1.3	0.0002
GO:0006414	translational elongation	7	1.56	0.0006
GO:0009084	glutamine family aa biosynthetic process	6	1.17	0.0007
GO:0019243	methylglyoxal catabolic process to D-lactate	3	0.32	0.0025
GO:0006635	fatty acid beta-oxidation	6	1.56	0.0036
GO:0006560	proline metabolic process	3	0.39	0.0047
GO:0006979	response to oxidative stress	23	11.86	0.0058
GO:0009627	systemic acquired resistance	6	1.75	0.0067
GO:0009644	response to high light intensity	7	2.33	0.0075
GO:0035195	gene silencing by miRNA	3	0.45	0.0078
GO:0009616	virus-induced gene silencing	3	0.45	0.0078
GO:0007021	tubulin complex assembly	3	0.45	0.0078
GO:0042128	nitrate assimilation	3	0.45	0.0078
GO:0009408	response to heat	13	6.22	0.0089
Mol. Function	**Term**	**Signif.**	**Exp.**	**P value**
GO:0016844	strictosidine synthase activity	5	0.93	0.0015
GO:0051087	chaperone binding	4	0.6	0.0018
GO:0004416	hydroxyacylglutathione hydrolase activity	3	0.33	0.0026
GO:0003725	double-stranded RNA binding	4	0.66	0.0029
GO:0009983	tyrosine aminopeptidase activity	2	0.13	0.0044
GO:0010385	double-stranded methylated DNA binding	2	0.13	0.0044
GO:0004356	glutamate-ammonia ligase activity	3	0.46	0.0083

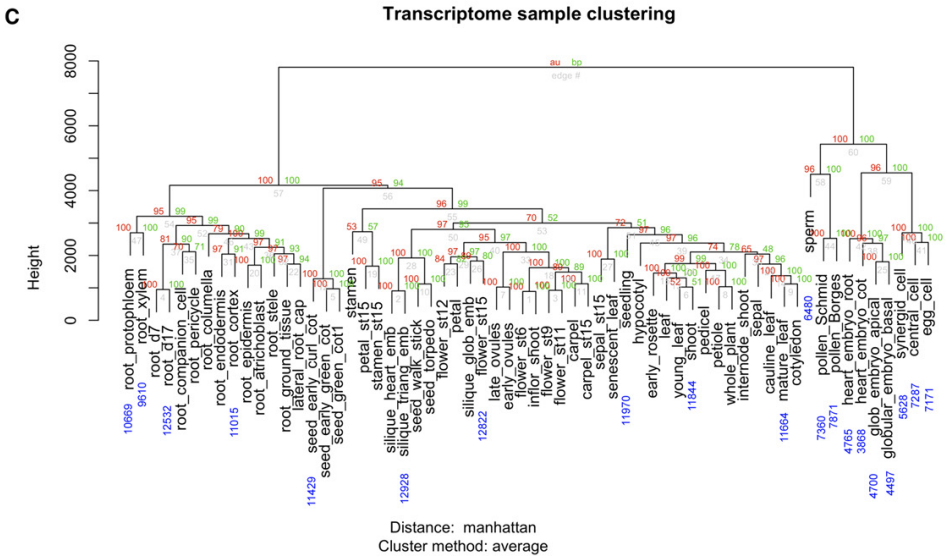


Figure 3. Female Gametophytic Cells Have Distinct Transcriptomes and Exhibit Differential Expression of Translational and Epigenetic Control Factors and Metabolic Pathways

(A) Heat map of differentially expressed genes identified by an analysis of variance (unadjusted $p < 0.01$) as a first screening method. Yellow denotes high expression, blue denotes low expression. Samples/genes were clustered via correlation distance and hierarchical agglomerative clustering. Colors are scaled per row. Note that most genes exhibit elevated expression in one cell type only, as opposed to elevated expression in two cell types.

(B) Gene ontology (GO) term enrichment table showing the significantly enriched biological processes and molecular functions among differentially expressed genes. The following abbreviations are used: signif., number of significant genes in a given term; exp., number of expected genes in a given term.

(C) Transcriptome clustering based on binary expression values (present/absent). The female gametophytic cells group together and are comparable in overall transcriptome sizes and/or compositions; female gametophyte, male gametophyte, and laser-captured embryo transriptomes form an outgroup to sporophytic tissues and cell types. Blue numbers denote overall transcriptome sizes (see also Figure S3 and Table S3).

egg. We found predominant expression of a subgroup of PAZ domain-encoding genes in the egg: among *DCL1*, *AGO1*, *AGO2*, and *AGO5*, the two functionally uncharacterized paralogs *AT5G21150* and *AT5G21030* (Figure 4A). These data suggest that small RNA pathways are a dominant feature of the generative female gamete of *Arabidopsis*. This could be important for protection against selfish genetic elements, as in the male gamete [30], or to regulate stem cell fate, paralleling epigenetic regulation in the germline of *Drosophila* and mammals (Figure 4B) [3]. In light of the recently reported genome-wide DNA hypomethylation [31, 32] and elevated production of siRNAs in the endosperm [9], we propose that

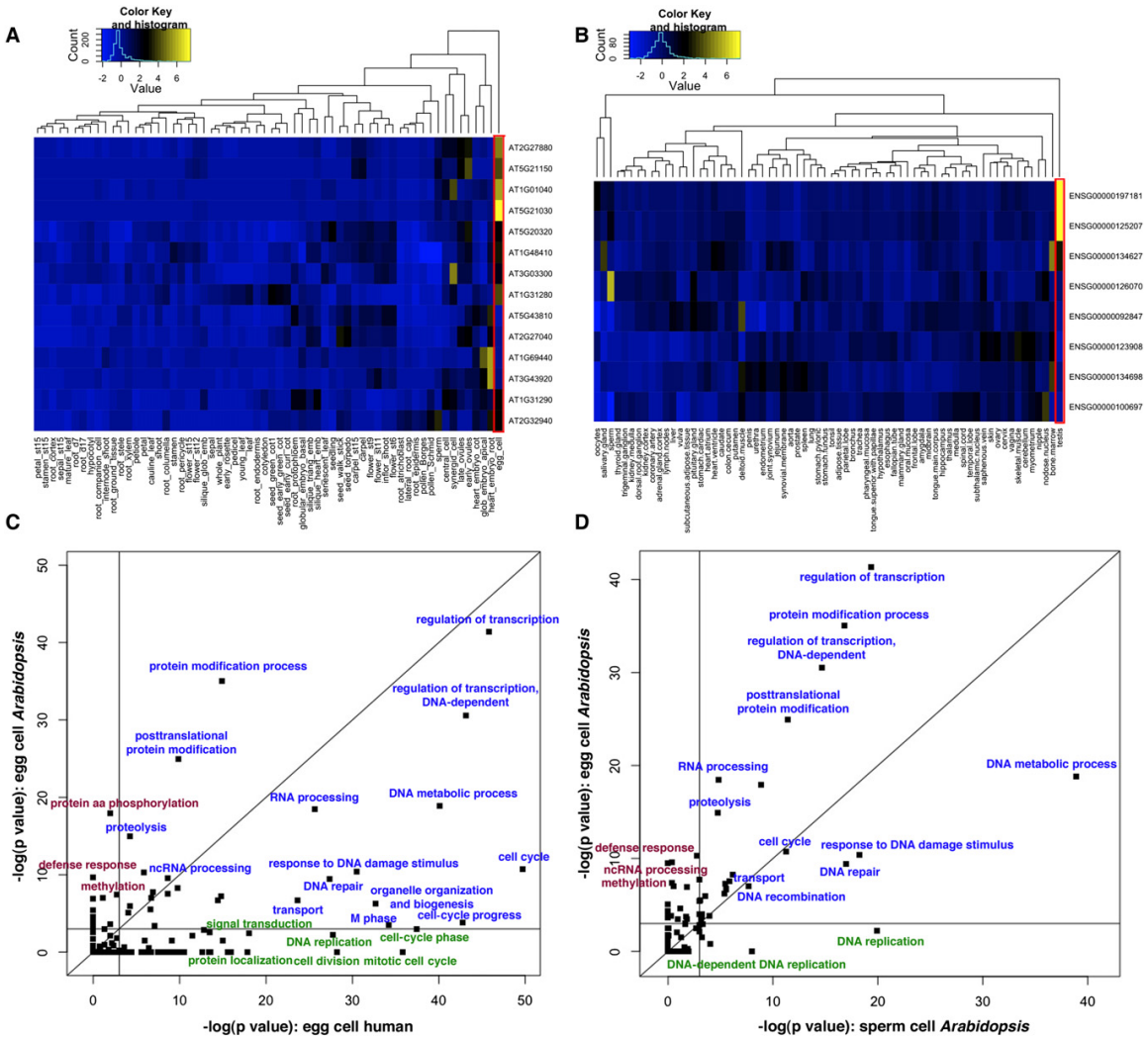


Figure 4. Gamete and Germ Cell Features in the *Arabidopsis* Egg Cell Transcriptome

(A and B) Expression of PAZ domain-encoding genes across the *Arabidopsis* and human tissue atlas. Yellow denotes high expression, blue denotes low expression. Genes/tissues are clustered via hierarchical agglomerative clustering (Euclidean distance), and signals are Z score normalized across rows.

(A) Heat map representation of mean expression across the *Arabidopsis* tissue atlas. Egg cells (red box), embryo tissues, and sperm exhibit elevated levels of expression of several genes encoding a PAZ domain when compared to the rest of the plant body.

(B) Heat map representation of mean expression across the human tissue atlas. Note that there is enriched expression of *Miwil2*, *Piwil1*, and *Piwil4* in testis (red box), as expected from their roles in germline development [3].

(C) Functional map of biological process terms comparing upregulation of GO functions in female gametes of *Arabidopsis* and humans. Negative logarithms of Bonferroni-adjusted p values from a Kolmogorov-Smirnov test for shifts toward higher signal values within a GO group are plotted. Vertical/horizontal lines indicate an adjusted p value of 0.001. Red and green denote terms that are significant in only one gamete; blue denotes terms that are significant in both. Selected data points are annotated in the figure; a full annotation can be found in Table S4.

(D) Functional map of biological process terms comparing *Arabidopsis* egg cells and sperm (as in C) (see also Figure S4 and Table S4).

embryo and endosperm development involves differential expression of epigenetic pathway components already established in the egg and central cell. Our data support a model in which siRNAs produced in the central cell and endosperm effect epigenetic gene regulation in the egg and embryo, respectively. This model could also explain egg-enriched expression of *RDM4* (Table S3), encoding a factor necessary for RNA-directed DNA methylation (RdDM) during development [33].

Sexual reproduction evolved in eukaryotes before the divergence of plants and animals. Thus, molecular aspects of gamete (syngamy) and nuclear fusion (karyogamy) may be conserved in the two lineages. Additionally, reproduction of angiosperms evolved several parallels to mammalian reproduction: (1) both lineages evolved anisogamy, (2) female gametes develop in a maternal environment providing nutrients, (3) mature gametes arrest prior to fertilization, (4) parental

imprinting evolved in both groups, and (5) selection based on male-male competition occurs in the prezygotic phase [34–36]. Thus, although it is generally not possible to compare plant and animal cell types, an interkingdom comparison of gamete transcriptomes may reveal basic molecular similarities. Therefore, we also constructed a tissue atlas of human transcriptomes, including oocytes and sperms (Supplemental Experimental Procedures), and compared expression signals within gametes against several tissues of the human body. We tested whether there is common up- or downregulation of orthologous pairs in human and *Arabidopsis* gametes but did not find a global trend (data not shown). However, from 7289 orthologous pair relations, we identified a total of 68 pairs with signals more than three standard deviations above the population mean in the eggs of both species. The latter are good candidates for genes that perform ancestral gamete functions, such as syngamy and karyogamy; however, little functional information is available for most of these genes (Table S4). The AtDRM1-HsDNMT3A orthology pair, encoding de novo DNA methyltransferases, is enriched in female gametes of both species. AtDRM1 is required for RdDM [37], providing a possible link between siRNA pathways and genome integrity maintenance that could function in both animal and plant female gametes.

We searched for overlaps of enriched functions in gametes of both lineages by comparing gene signal distributions within functional groups (GO groups or Pfam groups) across the respective tissue atlas. Functions enriched in human oocytes as detected by our analysis agreed with earlier studies (Table S4). When comparing globally enriched functions across the two species, we found overlaps of 26 "biological process" groups, 26 "molecular function" groups, and three Pfam domain groups (Figure 4A; Table S4), including RNA metabolism (transfer RNA and noncoding RNA processing, RNA polymerase activity), protein degradation, and cell-cycle control. The overlapping enrichment of the latter functional terms may indicate that, in both species, factors required for early cleavage cycles are deposited in the egg, which could be a consequence of the convergent evolution of anisogamy. That half of the female gametophytic mutants recovered to date show maternal effects [1, 17] supports the notion that, in plants, egg cells store cytoplasmic products as they do in animals [38].

Whether the functions and protein families we found enriched in the female gametes of both humans and *Arabidopsis* are indicative of conserved sexual elements or are a consequence of convergent evolution remains to be elucidated. However, 70% and 80% of the commonly enriched functional groups exhibit enrichment also in *Arabidopsis* and human sperm, respectively (Figure 4D; Figure S4). This could indicate that most represent ancestral gametic functions; however, evolutionary conclusions should await the availability of more gamete transcriptomes across different taxa. Comparisons among multiple species should allow a better dissection of gametic and gametophytic transcription modules within the female gametophyte. In addition, a better temporal resolution of female gametogenesis and early embryogenesis events will shed light on the molecular evolution of sexual processes and the transition between generations in plants and animals.

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures, four figures, and four tables and can be found with this article online at doi:10.1016/j.cub.2010.01.051.

Acknowledgments

We thank K. Byrne, K. Hokamp, M. Fares, K. Wolfe (Trinity College Dublin [TCD]), and G. Conant (TCD/University of Missouri-Columbia) for helpful discussions, M. Curtis (University of Zürich) for providing the pMDC162 cloning vector, T. Lehman and P. Grosscurt (University of Zürich) for access to the MMI-LAM system, C. Kägi (University of Zürich/University of Tübimgen) for testing various laser capture microscope systems, N. Kerk and T. Nelson (Yale University) for encouraging discussions, and T. Casneuf (University of Ghent/European Bioinformatics Institute Cambridge) for providing the negative probe set data. We are indebted to A. Patriagni and M. Künzli-Gontarczyk (Functional Genomics Center Zürich) for help with the array hybridizations. This work was supported by the University of Zürich, the Stiftung für wissenschaftliche Forschung through a grant of the Baumgarten-Stiftung (to U.G.), grants of the Science Foundation Ireland (to F.W.) and the Swiss National Science Foundation (to U.G.), and fellowships by the Netherlands Genomics Initiative (to K.V.) and the Deutsche Forschungsgemeinschaft (to A.S.).

Received: October 1, 2009
Revised: January 8, 2010
Accepted: January 12, 2010
Published online: March 11, 2010

References

1. Brukhin, V., Curtis, M., and Grossniklaus, U. (2005). The angiosperm female gametophyte: No longer the forgotten generation. Curr. Sci. *89*, 1844–1852.
2. Yadegari, R., and Drews, G.N. (2004). Female gametophyte development. Plant Cell *16* (*Suppl*), S133–S141.
3. Klattenhoff, C., and Theurkauf, W. (2008). Biogenesis and germline functions of piRNAs. Development *135*, 3–9.
4. Casson, S., Spencer, M., Walker, K., and Lindsey, K. (2005). Laser capture microdissection for the analysis of gene expression during embryogenesis of *Arabidopsis*. Plant J. *42*, 111–123.
5. Borges, F., Gomes, G., Gardner, R., Moreno, N., McCormick, S., Feijó, J.A., and Becker, J.D. (2008). Comparative transcriptomics of *Arabidopsis* sperm cells. Plant Physiol. *148*, 1168–1181.
6. Steffen, J.G., Kang, I.H., Macfarlane, J., and Drews, G.N. (2007). Identification of genes expressed in the *Arabidopsis* female gametophyte. Plant J. *51*, 281–292.
7. Pagnussat, G.C., Alandete-Saez, M., Bowman, J.L., and Sundaresan, V. (2009). Auxin-dependent patterning and gamete specification in the *Arabidopsis* female gametophyte. Science *324*, 1684–1689.
8. Kumakura, N., Takeda, A., Fujioka, Y., Motose, H., Takano, R., and Watanabe, Y. (2009). SGS3 and RDR6 interact and colocalize in cytoplasmic SGS3/RDR6-bodies. FEBS Lett. *583*, 1261–1266.
9. Mosher, R.A., Melnyk, C.W., Kelly, K.A., Dunn, R.M., Studholme, D.J., and Baulcombe, D.C. (2009). Uniparental expression of PolIV-dependent siRNAs in developing endosperm of *Arabidopsis*. Nature *460*, 283–286.
10. Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of *Arabidopsis thaliana* development. Nat. Genet. *37*, 501–506.
11. Brady, S.M., Orlando, D.A., Lee, J.Y., Wang, J.Y., Koch, J., Dinneny, J.R., Mace, D., Ohler, U., and Benfey, P.N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. Science *318*, 801–806.
12. Nishiyama, T., Fujita, T., Shin-I, T., Seki, M., Nishide, H., Uchiyama, I., Kamiya, A., Carninci, P., Hayashizaki, Y., Shinozaki, K., et al. (2003). Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: Implication for land plant evolution. Proc. Natl. Acad. Sci. USA *100*, 8007–8012.
13. Kasahara, R.D., Portereiko, M.F., Sandaklie-Nikolova, L., Rabiger, D.S., and Drews, G.N. (2005). *MYB98* is required for pollen tube guidance and synergid cell differentiation in *Arabidopsis*. Plant Cell *17*, 2981–2992.
14. Luo, M., Bilodeau, P., Dennis, E.S., Peacock, W.J., and Chaudhury, A. (2000). Expression and parent-of-origin effects for *FIS2, MEA*, and *FIE* in the endosperm and embryo of developing *Arabidopsis* seeds. Proc. Natl. Acad. Sci. USA *97*, 10637–10642.
15. Choi, Y., Gehring, M., Johnson, L., Hannon, M., Harada, J.J., Goldberg, R.B., Jacobsen, S.E., and Fischer, R.L. (2002). DEMETER, a DNA

glycosylase domain protein, is required for endosperm gene imprinting and seed viability in *Arabidopsis*. Cell *110*, 33–42.
16. Hejátko, J., Pernisová, M., Eneva, T., Palme, K., and Brzobohatý, B. (2003). The putative sensor histidine kinase CKI1 is involved in female gametophyte development in *Arabidopsis*. Mol. Genet. Genomics *269*, 443–453.
17. Pagnussat, G.C., Yu, H.J., Ngo, Q.A., Rajani, S., Mayalagu, S., Johnson, C.S., Capron, A., Xie, L.F., Ye, D., and Sundaresan, V. (2005). Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. Development *132*, 603–614.
18. Okuda, S., Tsutsui, H., Shiina, K., Sprunck, S., Takeuchi, H., Yui, R., Kasahara, R.D., Hamamura, Y., Mizukami, A., Susaki, D., et al. (2009). Defensin-like polypeptide LUREs are pollen tube attractants secreted from synergid cells. Nature *458*, 357–361.
19. Márton, M.L., Cordts, S., Broadhvest, J., and Dresselhaus, T. (2005). Micropylar pollen tube guidance by *egg apparatus 1* of maize. Science *307*, 573–576.
20. Silverstein, K.A., Graham, M.A., Paape, T.D., and VandenBosch, K.A. (2005). Genome organization of more than 300 defensin-like genes in *Arabidopsis*. Plant Physiol. *138*, 600–610.
21. Chen, Y.H., Li, H.J., Shi, D.Q., Yuan, L., Liu, J., Sreenivasan, R., Baskar, R., Grossniklaus, U., and Yang, W.C. (2007). The central cell plays a critical role in pollen tube guidance in *Arabidopsis*. Plant Cell *19*, 3563–3577.
22. Jones-Rhoades, M.W., Borevitz, J.O., and Preuss, D. (2007). Genome-wide expression profiling of the Arabidopsis female gametophyte identifies families of small, secreted proteins. PLoS Genet. *3*, 1848–1861.
23. Portereiko, M.F., Lloyd, A., Steffen, J.G., Punwani, J.A., Otsuga, D., and Drews, G.N. (2006). *AGL80* is required for central cell and endosperm development in *Arabidopsis*. Plant Cell *18*, 1862–1872.
24. Steffen, J.G., Kang, I.H., Portereiko, M.F., Lloyd, A., and Drews, G.N. (2008). AGL61 interacts with AGL80 and is required for central cell development in *Arabidopsis*. Plant Physiol. *148*, 259–268.
25. Franco-Zorrilla, J.M., Cubas, P., Jarillo, J.A., Fernández-Calvín, B., Salinas, J., and Martínez-Zapater, J.M. (2002). *AtREM1*, a member of a new family of B3 domain-containing genes, is preferentially expressed in reproductive meristems. Plant Physiol. *128*, 418–427.
26. Colombo, M., Masiero, S., Vanzulli, S., Lardelli, P., Kater, M.M., and Colombo, L. (2008). *AGL23*, a type I MADS-box gene that controls female gametophyte and embryo development in *Arabidopsis*. Plant J. *54*, 1037–1048.
27. Bemer, M., Wolters-Arts, M., Grossniklaus, U., and Angenent, G.C. (2008). The MADS domain protein DIANA acts together with AGAMOUS-LIKE80 to specify the central cell in *Arabidopsis* ovules. Plant Cell *20*, 2088–2101.
28. Kang, I.H., Steffen, J.G., Portereiko, M.F., Lloyd, A., and Drews, G.N. (2008). The AGL62 MADS domain protein regulates cellularization during endosperm development in *Arabidopsis*. Plant Cell *20*, 635–647.
29. Nam, J., Kim, J., Lee, S., An, G., Ma, H., and Nei, M. (2004). Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc. Natl. Acad. Sci. USA *101*, 1910–1915.
30. Slotkin, R.K., Vaughn, M., Borges, F., Tanurdzić, M., Becker, J.D., Feijó, J.A., and Martienssen, R.A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell *136*, 461–472.
31. Gehring, M., Bubb, K.L., and Henikoff, S. (2009). Extensive demethylation of repetitive elements during seed development underlies gene imprinting. Science *324*, 1447–1451.
32. Hsieh, T.F., Ibarra, C.A., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R.L., and Zilberman, D. (2009). Genome-wide demethylation of *Arabidopsis* endosperm. Science *324*, 1451–1454.
33. He, X.J., Hsu, Y.F., Zhu, S., Liu, H.L., Pontes, O., Zhu, J., Cui, X., Wang, C.S., and Zhu, J.K. (2009). A conserved transcriptional regulator is required for RNA-directed DNA methylation and plant development. Genes Dev. *23*, 2717–2722.
34. Bernasconi, G., Ashman, T.L., Birkhead, T.R., Bishop, J.D., Grossniklaus, U., Kubli, E., Marshall, D.L., Schmid, B., Skogsmyr, I., Snook, R.R., et al. (2004). Evolutionary ecology of the prezygotic stage. Science *303*, 971–975.
35. Marton, M.L., and Dresselhaus, T. (2008). A comparison of early molecular fertilization mechanisms in animals and flowering plants. Sex. Plant Reprod. *21*, 37–52.
36. Randerson, J.P., and Hurst, L.D. (2001). A comparative test of a theory for the evolution of anisogamy. Proc Biol Sci *268*, 879–884.
37. Cao, X., Aufsatz, W., Zilberman, D., Mette, M.F., Huang, M.S., Matzke, M., and Jacobsen, S.E. (2003). Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation. Curr. Biol. *13*, 2212–2217.
38. Baroux, C., Autran, D., Gillmor, C.S., Grimanelli, D., and Grossniklaus, U. (2008). The maternal to zygotic transition in animals and plants. Cold Spring Harb. Symp. Quant. Biol. *73*, 89–100.